

Review 31: Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere

פינט הסקרים:

המלצת קריאה מאלכסנדר וממייק: מומלץ מאוד לחובבי למידה self-supervised ולכל מי שמתעניין בלמידת ייצוג

בahirot_cetiha: בינוי פלוט

ידע מוקדם:

- יסודות תורת ההסתברות ותורת המידה
- יסודות של למידה self-supervised ולמידה ניגודית

ישומים פרקטיים:

- ניתן להשתמש בגישה המוצעת להפקה של ייצוג דאטה טוב יותר מאשר בגישה הקודמות של למידה ניגודית.

פרטי מאמר:

מאמר: [כאן](#)

קוד: [כאן](#)

הורסם בתאריך: ארכיב, 07.11.2020

הוזג בכנס: ICML, 2020

תחומי מאמר:

- למידה ייצוגית (representation learning)
- למידה ניגודית (contrastive learning)

כליים מתמטיים, מושגים וסימונים:

לוֹס נִיגוּדי (Contrastive Loss)	-
Radial Basic Function Kernel	-
עקרון InfoMax	-
הטכניקות החלשה (weak convergence) של מידות הסתברות	-

מבוא:

למידת ייצוג היא מונח גג למגוון שיטות המאפשרות לנו לבנות ייצוג נתוניים עצמאיים שניצן לנצלם למשימות downstream מגוונות. דוגמה מצינית ללמידה מודפסת ייצוג היא word2vec - אלגוריתם supervised שהוצע ב-2013 על ידי Tomáš Mikolov ועמיתו המשמש בגישה הלמידה הניגודית. האלגוריתם בונה וקטורים של ייצוגי מילים (אמבדינגים) בעלי מאפיינים רצויים מסוימים, למשל מילים בעלות משמעות דומה ממופות לנקודות (וקטורים) הקרובות במרחב הייצוג (embedding space). תכונה זו של ייצוג דاطה נקרא *ישור* (alignment). מודלי טרנספורמרים מרחיבים את יכולות של ייצוג word2vec והופכים אותן לתליים בהקשר. ככלומר ייצוג של מילה נתונה תליי במילים הקרובים אליה בטקסט. תכונה זו שדרגה את יכולות של מודלי שפה המבוססות על הטרנספורמרים אולם באותו הזמן וקטורי הייצוג שנוצרו באמצעות הטרנספורמרים סובלות צפיפות מאוד לא אחידה במרחב הייצוג ככלומר וקטורי הייצוג נוטים להתרכם באחור צר של מרחב הייצוג. תכונה זו עלולה לגרום למשל דנקרא "קורלציות בדווית" ([Mimno & Thompson, 2017](#); [Ethayarajh, 2019](#)) ככלומר קרבה בלתי רציה בין ייצוגים של מילים לא קשרות הפוגע בביטויים של המודל. צפיפות ייצוגים לא אחידה מהוות בעיה גם בדומיננטים אחרים כמו למשל הדומין הייזואלי. המאמר הנסקר מנסה לתת מענה לסוגיה זו.

תמצית מאמר:

המאמר מראה כי ייצוג דاطה, המופקים באמצעות מודלים שאומנו עם הלוֹס הניגודי (contrastive loss), עשויים להיות מאפיינים בשתי התכונות הטובות שהזכרנו קודם: אחידות וישור (uniformity). לאחר מכן, הם מציעים פונקציית לוֹס המשפרת את המאפיינים הללו ולבסוף הם מראים שפונקציית לוֹס זו יכולה להוביל לייצוגים טובים יותר מלה שהושגו באמצעות פונקציות לוֹס ניגודיות מסורתיות.

הסבר על הרעיון העיקרי:

למידה ניגודית היא אחת השיטות הנפוצות ביותר לבניית ייצוג דاطה (בדרכו כלל במרחב בעל ממד נמוך הנקרא לעיתים מרחב הילטנטני) עבור דאטassetים לא מתאימים. הנחתהasisון מאחרי טכניקה זו היא שלדוגמאות דומות יש וקטורי ייצוגים קרובים, בעוד שלדוגמאות לא דומות יש וקטורי ייצוג מרוחקים. בפרט, ברוב שיטות הלמידה הניגודיות נבנים זוגות דוגמאות דומות (חיבויות) וזוגות מדגם לא דומים (שליליים) במהלך האימון. מטרת הלמידה הניגודית מנסה בדרך כלל למקסם את היחס בין המרחקים בין ייצוגי זוגות חיובים ושליליים. בישומי ראייה ממחושבת למשל, שני קרופים שונים של אותה תמונה יוצרים זוג חיובי בעוד שקרופים מתומות שנבחרו באקראי יוצרים זוג שלילי.

כאמור במאמר זה, המחברים בוחנים את המאפיינים של ייצוג דاطה שאומנו באמצעות שיטות למידה המשמשים בלבד הניגוד. המאמר מראה כי ייצוג דاطה המופקים במהלך הלמידה הניגודית יש שתי תכונות הבאות:

1. **ישור (alignment)**: קרבה בין ייצוגים של של פיסות DATA קרובות
2. **אחדות (uniformity)**: ייצוג DATA מפולגים באופן אחיד בהיפר-ספירה ברדיו 1 (ראה הערה למטה). באופן אינטואיטיבי, אחדות של התפלגות היצוגים למרחב הלטני מצביעה על כך שהיצוגים "שומרים" **כמויות מקסימלית של מידע** של הדטה המקורי.

הערה: המאמר מוסיף אילוץ של נורמה יחידה על ייצוג DATA. מספר עבודות קודמות מצאו כי אילוץ זה תורם לשיפור יציבות של תהליכי האימון של שיטות למידה ניגודיות. נראהשהסתיבת לכך טמונה בשימוש "כבד" במफולות פנימיות בפונקציות לוס של שיטות אלו. נציין כי למיטב ידיעתנו לא קיימת הוכחה ריגורוזית לכך שנורמה יחידה מהוות תכונה "מעילה" עבור ייצוג DATA.

התוצאה העיקרית של המאמר קובעת כי הלמידה הניגודית ממקסמת את שני המאפיינים שהוזכרו לעיל כאשר מספר הדגימות השיליות שואף לאינסוף. במליל פשטוט, כאשר מספר הדוגמאות השיליות בミニ-באטץ גבוהה, אופטימיזציה של פונקציית הלוס הניגודית מובילה לייצוג DATA מושרים ומפולגים באופן אחיד.

כמה עבודות ציינו כי הגדלת מספר הדוגמאות השיליות בשיטות למידה ניגודית תורמת לשיפור של ייצוג DATA. מנוקوت מבט זו, החלפת פונקציית מטריה של למידה ניגודית בכו שמאਪטמת אחדות וישור היצוגים באופן ישיר עשויה להוביל לייצוג DATA חזקים יותר. על מנת להטיל את אילוצי האחדות וישור על וקטורי ייצוג DATA, המחברים הציגו מדדים, מעוגנים תיאורטיות, למידית היישור והאחדות של היצוגים. לבסוף המאמר הראה כי שילוב של הלוס המוצע (ישור ואחדות) יחד עם הלוס הניגודי, הצלח ליצור ייצוג DATA טובים יותר.¹

התמונה למטה ממחישה את תכונות היישור והאחדות החזקות של ייצוג DATA שנלמדו באמצעות למידה ניגודית (2 תמונות משמאלי בתחתית) על DATAsets ללא תיוגים. מענין לציין כי ייצוג DATA שנלמדו באמצעות למידה supervised (DATA מתויג) מגינות גם כן רמה גבוהה של אחדות וישור (2 תמונות משמאלי ביותר בשורה האמצעית).

¹ איקות ייצוג DATA נקבעת לרוב על ידי מידת ההפרדה הליניארית של קלאסטרים המורכבים על ידי דוגמאות מקטגוריות שונות. קלאסטרים מופרדים היטב מוצביהם על כך שהיצוגים הנלמדים תפסו את התוון הסמנטי של הנתונים.

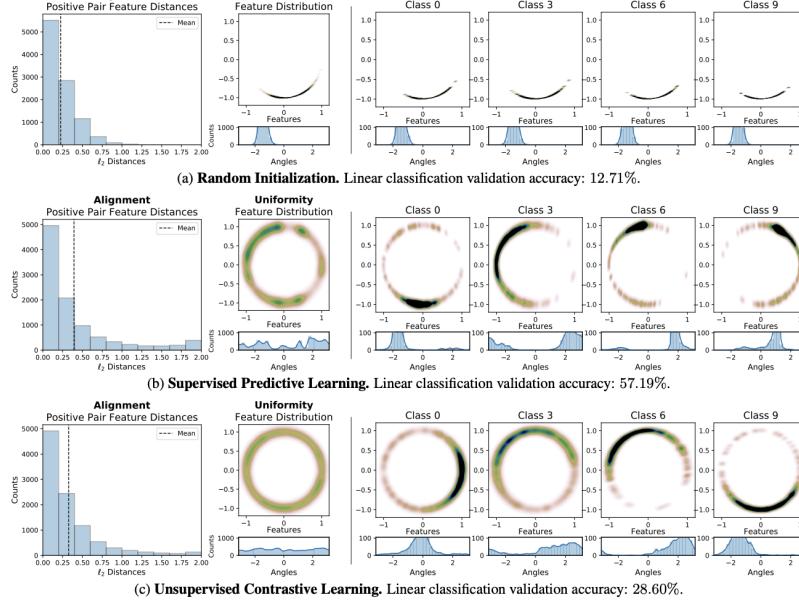


Figure 3: Representations of CIFAR-10 validation set on S^1 . **Alignment analysis:** We show distribution of distance between features of positive pairs (two random augmentations). **Uniformity analysis:** We plot feature distributions with Gaussian kernel density estimation (KDE) in \mathbb{R}^2 and von Mises-Fisher (vMF) KDE on angles (i.e., $\arctan 2(y, x)$ for each point $(x, y) \in S^1$). **Four rightmost plots** visualize feature distributions of selected specific classes. Representation from contrastive learning is both *aligned* (having low positive pair feature distances) and *uniform* (evenly distributed on S^1).

פינת האינטואיציה:

באו ננסה לספק קצט תובנות לגבי מושע יישור ואחדות יכולם להיות מאפיינים טבעיות של ייצוג דатаה, שהופקו באמצעות למידה עם פונקציית הלוי הניגודית. קודם כל נתחיל מறען מהי פונקציית הלוי הניגודית (ה策ורה הנפוצה ביותר):

$$\mathcal{L}_{\text{contrastive}}(f; \tau, M) \triangleq \mathbb{E}_{\substack{(x, y) \sim p_{\text{pos}} \\ \{x_i^-\}_{i=1}^M \sim p_{\text{data}}}} \left[-\log \frac{e^{f(x)^\top f(y)/\tau}}{e^{f(x)^\top f(y)/\tau} + \sum_i e^{f(x_i^-)^\top f(y)/\tau}} \right], \quad (1)$$

נניח שלדוגמאות חיוביות יש את אותו ייצוג / מישرات בצורה מושלמת. נציין כי במקרה זה המכפלת הפנימית שלהם תהיה שווה ל-1, מכיוון נורמה של כל ייצוג שווה ל-1. לכן הביטוי קודם מקבל את הצורה הבאה:

$$\mathbb{E}_{\substack{x \sim p_{\text{data}} \\ \{x_i^-\}_{i=1}^M \sim p_{\text{data}}}} \left[\log \left(e^{1/\tau} + \sum_i e^{f(x_i^-)^\top f(x)/\tau} \right) \right],$$

כאשר מספר הדוגמאות השליליות M גדול מאוד, המינימום של הביטוי האחרון מושג כאשר המרחקים בין זוגות של ייצוגים הוא מקסימלי (המכפלה הפנימית באקספוננט קרובה לאפס ככל האפשר). אז יישור ואחדות נראים כאמור טבעיות של ייצוג דата המשיג ערך קטן של הלוס הניגודי.

המשפט העיקרי:

כעת נדון במשפט הראשי של המאמר:

Theorem 1 (Asymptotics of $\mathcal{L}_{\text{contrastive}}$). *For fixed $\tau > 0$, as the number of negative samples $M \rightarrow \infty$, the (normalized) contrastive loss converges to*

$$\begin{aligned} \lim_{M \rightarrow \infty} \mathcal{L}_{\text{contrastive}}(f; \tau, M) - \log M = \\ -\frac{1}{\tau} \mathbb{E}_{(x, y) \sim p_{\text{pos}}} [f(x)^T f(y)] \\ + \mathbb{E}_{x \sim p_{\text{data}}} \left[\log \mathbb{E}_{x^- \sim p_{\text{data}}} \left[e^{f(x^-)^T f(x)/\tau} \right] \right]. \end{aligned} \quad (2)$$

We have the following results:

1. The first term is minimized iff f is perfectly aligned.
2. If perfectly uniform encoders exist, they form the exact minimizers of the second term.
3. For the convergence in Equation (2), the absolute deviation from the limit decays in $\mathcal{O}(M^{-1/2})$.

The theorem states that the contrastive loss is equal to the sum of 2 terms for the number of negative examples M approaching infinity (bit simplified for clarity).

המשפט קובע כי הלוס הניגודי שווה לסכום של 2 איברים הבאים כאשר מספר הדוגמאות השליליות M שואף לאינסוף (מעט מופשט לצורך הבהירות):

איבר 1: ממוצע עבור ייצוג דата "המושרים" בצורה מושלמת (הייצוגים של כל הזוגות חיוביים זהים).

איבר 2: ממוצע כאשר הייצוגים מפולגים באופן אחד.

AIR לאמן ייצוגים מושרים ומפולגים אחד?

אחד נראה שהחלפת לוס ניגודי בלוס המשלב אחדות עם יישור, מובילה לייצוגי דата חזקים יותר. שאלת היא כיצד אנו מאמנים מודל מסוגל להפיק ייצוגים עם תכונות אלו? המחברים מציעים לאוסף יישור ואחדות על הייצוגים באמצעות פונקציות לוס הבאות:

עבור יישור: לוס היישור מוגדר בתור מרחוק ממוצע בין ייצוגים של זוגות חיוביים:

$$\mathcal{L}_{\text{align}}(f; \alpha) \triangleq \mathbb{E}_{(x, y) \sim p_{\text{pos}}} [\|f(x) - f(y)\|_2^\alpha], \quad \alpha > 0.$$

עבור איחידות: על מנת לאכוף איחידות על "יצוגי DATA", המאמר משתמש במה שמכונה פונקציה רדיאלית גאותית (הידועה גם כפונקציית בסיס רדיאלית או RBF). עם RBF, המרחק בין וקטורי ייצוג x ו- y מוגדר בצורה הבאה:

$$G_t(u, v) \triangleq e^{-t\|u-v\|_2^2} = e^{2t \cdot u^\top v - 2t}, \quad t > 0,$$

נציין כי השווין השני נובע מהנורמה היחידה של וקטורי ייצוג x ו- y . אפשרות פונקציה לוס "האוכפת" איחידות מוגדרת בתור:

$$\begin{aligned} \mathcal{L}_{\text{uniform}}(f; t) &\triangleq \log \mathbb{E}_{\substack{x, y \sim \text{i.i.d.} \\ p_{\text{data}}}} [G_t(u, v)] \\ &= \log \mathbb{E}_{\substack{x, y \sim \text{i.i.d.} \\ p_{\text{data}}}} \left[e^{-t\|f(x)-f(y)\|_2^2} \right], \quad t > 0. \end{aligned}$$

אבל למה שימוש בפונקציה לוס זו "תפקיד" לנו "יצוגים מפולגים באופן יחיד"? מסתבר (זהה אמר מוכיח זאת) כי פונקציית הלוס מבוססת RBF **ממודעתה** כאשר וקטורי **היצוג המפולגים באופן אחד על היפר-ספרה** בעלת **רדיוויס 1**. בפשטות, עבור גודל מדגם (DATASET) גדול מאוד, וקטורי ייצוג המודיעים את פונקציית לוס זו "יכסו את פני השטח של היפר-ספרה היחידה באופן כמעט לגמרי".

הערה: אין פונקציית לוס האוכפת איחידות והן זו האוכפת היישור של "יצוגים הינם פחות מבחינה חשובה מאשר הלוס הניגודי הסטנדרטי עקב היעדר פעולה softmax בו".

לכן אימון רשת נירונית עם שילוב של פונקציות לוס הניל נראה דרך סבירה להציג "יצוגי DATA" ומושרים ומפולגים באופן יחיד.

הישגי מאמרה:

המחברים מצאו כי פונקציית לוס המוצעת (איחידות + יישור) וגם שילובה עם הלוס הניגודי הסטנדרטי הצליח להפיק "יצוגים חזקים יותר", וכתוואה מכך השיג ביצועים טובים יותר במספר שימושות downstream (סיווג ואומדן عمוק) במספר DATASETS (כולל [NYU Depth V2](#), [ImageNet](#), [ImageNet100](#), [BookCorpus](#)).

ג.ב.

במשך זמן רב, יישור ואיחידות הוכרו כמאפיינים טובים של "יצוגי DATA". המאמר הצליח להצביע על קשרים בין מאפיינים אלה לבין מספר שיטות אימון של למידה ניגודית.

Review 32: GAN-Control: Explicitly Controllable GANs

פינט הסקור:

המלצת קרייה ממיק: חובה לאוהביGANים.

בahirot_citing: טוביה מאד.

רמת היכרות עם כלים מתמטיים וטכניקות של DL/ML הנדרשים להבנת מאמר: נדרש הבנה טובה בארכיטקטורותicas של הגאנים (StyleGAN2) וידע בסיסי בנושא אימון של הגאנים. בנוסף נדרש הבנה בסיסית של עקרונות הלמידה הניגודית.

ישומים פרקטיים אפשריים: יצרה של תמונות פוטוריאלי-סטיות בעלות מכלול מוגדר של פיצרים ויזואליים כגון גיל, תנחת ראש, צבע שיער וכדומה בכמה דומיינים כמו תמונות פנים מציאות ותמונות פרצופים של חיות.

פרטי מאמר:

lienek למאמר: זמן-can

lienek לקוד: לא שותף בארכיב

פורסם בתאריך: 07.01.21, בארכיב

הציג בכנס: לא ידוע

תחומי מאמר:

- גאנים (GANs).

כלים מתמטיים, טכניקות, מושגים וסימונים:

- למידה ניגודית (contrastive learning)
- אימון של גאן עם פיצרים מופרדים בצורה מפוזרת.



Figure 1: We propose a framework for training GANs in a disentangled manner which allows for explicit control over generation attributes. Our method is applicable to diverse controls in various domains. First row (left to right) demonstrates our control over facial expression, age and illumination of human portraits. Second row (left to right) demonstrates our control over artistic style, age and pose of paintings. Third row demonstrates our pose control over faces of dogs.

מבוא והסבר כללי על תחום המאמר:

ליצירה של תמונות פוטוריאלייטיות בהינתן פרמטרים ויזואלים (כמו גיל, צבע שער, תוארה וכדומה) באיכות גבוהה יש שימושים רבים במגוון תחומים כגון עיצוב גרפי, משאקי וידאו, קולנוע, תחום הצלומים הרפואיים ועוד. בשנים האחרונות נרשמו כמה בפריצות דרך בתחום זהה במיחודה בפיתוח מודלים יצירת צילומי פנים (face images) בעלות מכלול מוגדר של פיצרים ויזואליים. מודלים אלו בדרך כלל משתמשים בשיטות מידול 3D וסובלים לרוב מעליות יצרה גבוהות ובעלות שונות נמוכה (יצירות תמונות דומות אחת לשניה). מצד שני מודלי גאנים עכשוויים כמו StyleGAN2 מפגינים יכולת מרשיםה ביצירת תמונות פוטוריאלייטיות באיכות מאוד גבוהה ובעלות יצרה סבירה אך מתאפשרות ליצור תמונות בעלות פיצרים ויזואליים נתוניים. יש עבודות המשלבות את שתי הגישות האלה ומצליחות ליצור תמונות פנים פוטוריאלייטיות באיכות מרשימה בעלות תכונות כמו תנוחה, תוארה וסוג הבעת פנים. אך מכיוון שמודלים אלו מtabssים על מידול 3D הם לא מאפשרים ליצור, למשל, תמונה של אדם בגיל מסוים כי מודלי 3D אינם מאפשרים זאת. בנוסף קשה להרחיב שיטות אלו לדומינינט קרובים כמו תמונות מציאות או צילומים של פרצופי חיות אם אין ברשותנו מודלים של 3D המתאימים לתחומים אלו.

המאמר הנסקר מציע גישה, הנקראת GAN-Control, הנותנת מענה לחולשה זה ומציע שיטה המאפשרת ליצור תמונות בעלות תכונות ויזואליות מוגדרות בצורה מפורשת. השיטה מצילהה ליצור תמונות פוטוריאלייטיות ב 3 דומינינטים שונים: צילומי פנים, תמונות מציאות וצלומי פרצופים של חיים. הגישה שלהם מאפשרת ליצור תמונות מגוונות לגיל, תוארה, תנוחה וצבע שער נתונים בצורה מפורשת לתחומים הנ"ל.

השיטה המוצעת מורכבת משני שלבים עיקריים:

- אימון של גן עם פיצרים מופרדים בצורה מפורשת(explicitly disentangled features): בגדול (הסביר המפורט ינתן בפרק הבא) מחלקים את המרכיב הלטנטי הראשון Z (ראה הערה על המרכיבים הלטנטיים למטה) לחת-MRI Z אשר כל תת-MRI אחראי על פיצר ויזואלי ספציפי של התמונה המגונרטת (חת-MRI האחרון אינו אחראי על פיצר ספציפי ואחראי על שאר הפיצרים הלא בשלטים של תמונה)
- אימון של רשת הממפה פיצרים ויזואליים נתונים (בצורה מפורשת כמו למשל גיל או זווית צילום של תמונות פנים) לחת-MRI הלטנטי "שלה". כתוצאה לכך תת-MRI age_Z , למשל, האחראי על הגילאים יחולק ל"איזורים" כך שכל "איזור" אחראי על גיל מסוים. שלב זה מאפשר לגונרטת תמונה בעלת פיצרים ויזואליים נתונים:

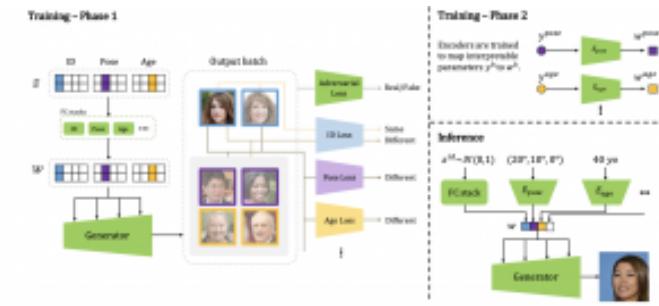


Figure 2: **Explicitly controllable GAN:** In Phase 1, we construct every batch so that for each attribute, there is a pair of latent vectors sharing a corresponding sub-vector, z^k . In addition to the adversarial loss, each image in the batch is compared in a contrastive manner, attribute-by-attribute, to all others, taking into account if it has the same or a different sub-vector. In Phase 2, encoders are trained to map interpretable parameters to suitable latent vectors. Inference: An explicit control over the attribute k is achieved by setting the E_k 's input to a required value.

הערה לגבי המרכיבים הלטנטיים של אולו: GAN-Control

המאמר משתמש בארכיטקטורה של StyleGAN2 שישי לו שני מרחבים לטנטים Z ו- W כאשר הראשון הינו מרחב הלטנטי "הגולמי" הסטנדרטי של הגאים והשני W הינו מרחב הסגןון הלטנטי עם פיצרים מופרדים האחראים על אספקטים ויזואליים שונים של התמונה המוגנרטת. המיפוי מ- Z ל- W ממודל עי"ר רשת MLP בעלת 8 שכבות(עם משקלים נלמדים כמובן). המאמר מציע לממד כל מיפוי מתת-מרחבים z_i ל w_i עי"ר רשת MLP שלה.

הסבר של רעיונות בסיסיים:

נדון כתת איר מאנים את אולו GAN-Control לגנרט תמונות מפיצרים מופרדים.

שלב אימון 1: המאמר מציע להשתמש בלבד היגודי לכל פיצר. בשביל כך כל באטץ' גבנה מזגות של וקטורים לטנטים z המכילים תת-וקטור מסווקף b_k המתאים לתת-מרחב הלטנטי (Z_k) כאשר שאר התת-וקטורים הינם שונים. נציין שעבור זוג דוגמאות כללו, הגרנרטור של אולו GAN-Control אמר לו ליצור שתי תמונות דומות בפיצר $-k$ בלבד ובדוחות בשאר הפיצרים. בשביל לבנות הלאו היגודי המאמר מגדיר מרחק בין זוג תמונות 1 ו- 2 א' מבחינת הפיצר k , המסומן $(2_1)_k - (1_1)_k$ המודד דמיון בין התמונות בפיצר k . למשל עבור הפיצר שהוא זהות של אדם בתמונה k מודד "עד כמה 1_1 ו- 2_1 מתראות אותו אדם".

از איך בעצם מחשבים את הלאו היגודי כאן?

- עבור זוג תמונות עם פיצר k זהה, אנו מנסים להגדיל את הדמיון בין התמונות הנוצרות מבחינת פיצר k . •
כלומר המטריה הינה להקטין את המרחק ביןיהן, הניתן עי"ר D_k מוגדר כמקסימום בין הפרש של $(2_1)_k$ וקבוע c_k , לפחות אפס. ככלומר אנו "שואפים" שהמרחב המקסימלי בין 1_1 ו- 2_1 מבחינת פיצר k יהיה c_k לכל היותר.
- עבור זוג תמונות עם פיצר k שונה, המטריה למקסם את המרחק בין התמונות מבחינת פיצר זהה. הלאו במקורה זהה מוגדר עי"ר המקסימום בין אף סטי ההפרש בין קבוע $*c_k$ ו- $(2_1)_k - D_k$. בדומה לסייף הקודם המטריה כאן לא רצוי למרחק D_k להיות לפחות הפחות $*c_k$.
- פונקציית לוס מוגדרת כסכום של הלאוים על כל הפיצרים •

בסוף, מקובל בagan, מוסיפים להלאו היגודי את הלאו האדברסיאלי של StyleGAN2.

חישוב מרחק k_D: בשביל למדוד מרחק בין תמונות מבחינת פיצ'ר A המאמר משתמש ביצוג במילד נמור k_M של תמונה הנקבנה עי"ר רשות המאמנת למטרת זההו פיצ'ר k. למשל עבור דמיון בין תמונות מבחינת גיל, המאמר משתמש ביצוג מילד נמור שנבננה עי"ר הרשות להזיהוי גיל, למדידת דמיון בין תמונות מבחינת זהות של אדם המצלום, לוקחים את היצוג מהרשות להזיהוי פנים (face detection). המאמר משתמש במרקקי L2, L1 ומרקק cosine cosme למדידת מרחק בין וקטורי ייצוג של תמונות מבחינת פיצ'רים ויזואליים שונים.

שלב איתון 2: המאמר מציע לבנות מיפוי (אלמן רשות) נפרד לכל פיצ'ר A ויזואלי כמו גיל (20, 30, 40,...) או זווית צילום (0, 5, 15,...) למרחב "סגןון" לטנסי שלו k_W . אבל איך מאמנים את הרשותות אלו? המאמר מציע את דרך אלגנטית לעשות זאת:

- מרגלים מספר וקטורי Z.
- ממפים אותם למרחב "סגןון" W (זכרים שכבר אימנו את המיפוי הזה בשלב הראשון).
- מגנרטים תמונות מוקטורי הסגןון W ומעבירים את התמונות דרך הרשותות להזיהוי כל פיצ'ר A_y.
- בונים דאטහסט מזוגות (k_y , k_w) לכל פיצ'ר A.
- מאמנים רשותות מקודדות k_E המפות k_y ל- k_w (גם לכל פיצ'ר בנפרד).

از מה קורה בזמן האינפרנס? זה מאד פשוט – מזינים פיצ'רים ויזואליים k_y לרשותות מקודדות k_E ובונים וקטורי הסגןון k_w . בסוף משתמשים בגנרטור המאומן בשביל ליצור תמונה.

הישגי מאמר:

המחברים השוו את התמונות הנבנות עם GAN-Control עם אלו שנוצרו עם השיטות DFG ו- CONFIG (שיוצרות תמונות פנים עם פרמטרים ויזואליים נשלטים וմבוססות על גישות מידול 3D). המאמר הצליח להוכיח את עליונותה של GAN-Control על שיטות אלו ב- 3 דומיניניסונים בשני היבטים הבאים:

- **מרקק inception של פרשה (FID)** משופר(נמור יותר) המצביע בדרך כלל על תמונות יותר פוטוריאלי-סתיות.
- **דיק** משופר מבחינת התאמה של תמונה לפרמטרים הויזואליים שאיתם היא נבנתה. למשל עבור תמונה הנוצרת עם זווית צילום של 30 מעלות, GAN-Control הצליח ליצור תמונות עם זווית צילום קרובה יותר ל- 30 מעלות במעט ועם שונות נמוכה יותר (זווית צילום של תמונה נמדדת עי"ר רשות ייעודית מאומנת למשימה זו). דיק מבחינת פרמטרים אחרים נמדד לצורה דומה.

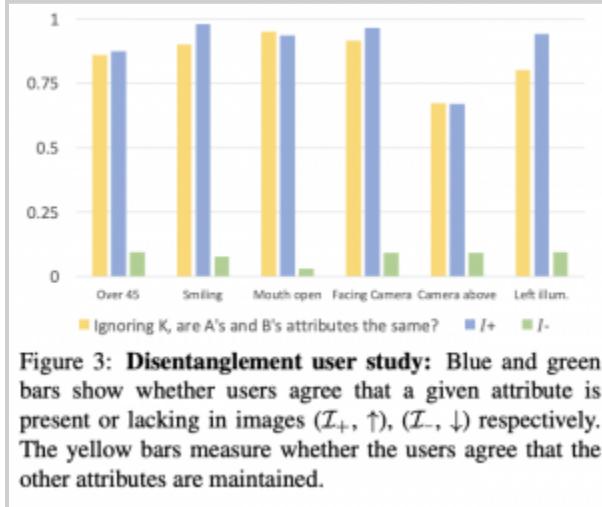


Figure 3: **Disentanglement user study:** Blue and green bars show whether users agree that a given attribute is present or lacking in images (\mathcal{I}_+ , \uparrow), (\mathcal{I}_- , \downarrow) respectively. The yellow bars measure whether the users agree that the other attributes are maintained.

GAN Version	Ours 512x512	DFG [13] 256x256	CONFIG [30] 256x256
Vanilla	3.32	5.49	33.41
Controlled	5.72	12.9	39.76

	Ours	DFG	CONFIG
Synthetic comparison	67%	22%	11%
Synthetic vs. real	47%	27%	16%

Table 2: **Photorealism user studies \uparrow :** (First row) users were asked to vote for the most realistic image from triplets of synthetic images (Ours, DFG, CONFIG). (Second row) users were shown pairs of images – one synthetic and one from the FFHQ dataset – and were asked to choose the real one from the two.

דאטאסתים: FFHQ, MetFaces, AFHQ

נ.ב. המאמר מציע גישה מאוד יעהה ואינטואיטיבית ליצירת תמונות בעלות פיצ'רים ויזואליים נשלטים באיכות גבוהה יותר מהשיטות המתחזרות. הגישה מסוגלת ליצור תמונות פוטו-ריאליסטיות בעלות תכונות ויזואליות נתונות ב-3 דומיניים שונים: צילומי פנים, צילומי פרצופי חיות ותמונות מציאות. הקוד לא שותף כרגע (אני מניח שהמחברים טרם הספיקו למלא בקשה פטנט – זה חשוב לחברת כמו AMAZON). אני די בטוח הקוד יפורסם ממש قريب. אני גם מכך להראות את השימושים של גישה זו בדומיניים נוספים.

Review 33: PreTrained Image Processing Transformer

פינת הסוקר:

המלצת קריאה ממוקם: רק עם קשה לכם להירדם בלילה (שווה לאלו שמתעניינים במשימות low-level בתחום עיבוד תמונה).

בהירות כתיבה: ביןוני מינום.

רמת היכרות עם כלים מתמטיים וטכניקות של DL/ML הנדרשים להבנת מאמר: היכרות עם מושגי יסוד של DL.

ישומים פרקטיים אפשריים: הגישה המוצעת במאמר יכולה לשמש כ启发ית אימון למשימות כמו סופר-ריזולוציה, ניקוי רעש רגיל או הסרת רעש גשם (deraining) עבור דאטאסטיטים קטנים.

פרטי מאמר:

לינק למאמר: [זמן להורדה](#).

לינק לקוד: לא הצלחתי לאתר.

פורסם בתאריך: 20.12.03, בארכיב.

הציג בכנס: לא מצאתי מידע על כר.

תחומי מאמר:

- למידה עם משימות מרובות (multi-task learning – MLT).
- למידה ניגודית (CL – contrastive learning).

כלים מתמטיים, טכניקות, מושגים וסימונים:

- טרנספורמר ויזואלי (הפועל על פאצ'ים של תמונות).
- LOSS ניגודי (contrastive loss).
- משימות low-level של הראייה הממוחשבת כמו סופר-ריזולוציה, ניקוי רעשים וכדומה.

لينקים להסבירים טובים על מושגי יסוד במאמר:

- [אימון מוקדים של רשותות](#).
- [טרנספורמר 1, טרנספורמר 2](#).
- [למידה ניגודית](#) (contrastive learning).

מבוא והסבר כללי על תחום המאמר:

רשותות נוירונים הפכו לכלי הנפוץ ביותר עבור מגוון רחב של משימות בראייה הממוחשבת החל ממשימות high-level כמו סיוג, סגמנטציה, זיהוי אובייקטים וכדומה וכלה במשימות low-level-low כמו ניקוי רעש, סופר-רזולוציה, שחזור חלקיים פגומים של תמונה (painting) ועוד. עקב דמיון בין משימות low-level-low-soו רבות ניתן לצפות שמודול (או הייצוג שנבנה באמצעותו) שאומן על דאטהסט מסוים יהיה שימושי גם עבור דאטהסטים אחרים. אז איך ניתן לנצל את הדמיון הזה? מאמנים מודול אחד על דאטהסט גדול (pretraining) ובדרך כלל (אך לא תמיד) על אותה משימה (!!!) ולאחר מכן מכילים את המודל המקורי (fine-tuning) על דאטהסט אחר (נקרא לו דאטהסט מטרה) שיכל להיות קטן בהרבה. די ברור שככל הדמיינים של הדאטהסטים דומים יותר, היעילות של האימון המקדמים עולה.

גישה זו טומנת בעצם שתי אתגרים עיקריים:

- לא תמיד יש דאטהסטים זמינים לאימון מקדמים (למשל בדומיין הרפואי או בדומיין של תמונות לוויין) לשימוש נתונה.
- לא תמיד ניתן לדעת לאיזו משימה יאומן מודול מודול בתהליך הcoil שמקשה על בחירה של דאטהסט לאימון מקדמים (נקרא לדאטהסט זה דאטהסט מקור).

תמצית מאמר:

במטרה להתמודד עם אתגרים אלו, המאמר מציע שיטת אימון מקדמים למשימות מרובות בדומיין הייזואלי. הם קראו לגישה שלהם IPT - Image Processing Transformer (כמו שאתם יכולים לנחש הארכיטקטורה שלהם מבוססת על הטרנספורמר). IPT מרכיב מרובה מרכיבים (רשתות עיקריות שמאומנים לכמה משימות במקביל (!!)):

- רשתות "ראשים" (heads): מספר ראשים שווה למספר משימות שעליין IPT מאומן, (ראש פר משימה). כל ראש הוא למעשה רשת קונולוציונית שיעדה להפיק מהקלט פיצרים רלוונטיים לשימוש שעלה אחריו הראש הזה.
- מקודד (encoder): הפלט של כל ראש מזון למקודד הסטנדרטי של הטרנספורמר.
- מפענה (decoder): הפלט של המקודד עובר למפענה די סטנדרטי של הטרנספורמר עם שניי קטן (יפורט בפרק הבא).
- רשתות זנבות (tails): מספר "זנבות" שווה למספר משימות (כמו ראשיים) והם מיועדים בשבייל ליצור קלט עבור כל שימוש (שעשו להיות במימד שונה לכל שימוש).

תוספת של האימון הניגודי (contrastive training) ל-IPT: נציין שקיים מגוון רחב של משימות בעלות אופיינים שונים ובדומיינים שונים, שלא ניתן לאמן את כולם במהלך אימון מקדמים. אז בשבייל לשפר את עצמת הייצוג של תמונה, המופק ע"י IPT, המאמר מציע לאמן אותן בשיטת הלמידה הניגודית (עם הלווי הניגודי הקלאסטי) בנוסף לאימון על מספר משימות low-level-low-so של עיבוד תמונה.

הסבר של רעיונות בסיסיים:

תחילה בואו נבין איך בונים קלט למוקוד. נתחיל מזה שהארQUITקטורה שלו זהה לזה של הטרנספורמר. הקלט למוקוד של הטרנספורמר הסטנדרטי (למשימות NLP) הוא האמבדינגס (embeddings) של טוקנים במשפט שמתווסף אליום קידוד מיקומי (positional encoding) שמטרתו "להעביר" למוקוד את המיקום של המילה במספט. ב- IPT עושים משהו דומה רק שבמקרים טוקנים יש לנו אטז'ים של תמונה (בגודל של 48x48). נציג שלhalbידל מהטרנספורמר הקלאסטי, הקידודים המיקומיים כאן הינם נלמדים (זה מה שעשו במאמר המפורטם [An](#)

self-attention Image is Worth 16×16 Words
אתה ולא 2.

הפלט של מקודד ונכנס למפענה שהוא מאד דומה לזה של הטרנספורמר המקורי עם שני הבדלים. הבדל הראשון הוא בנוסף לפלט של המקודד גם קידודים מיקומיים נלמדים פרט לשינה (!! מזינים למפענה (הסכם שלהם). המחברים טוענים כי תוספת זו תרמה רבות לביצועו המודל. ההבדל השני הוא העדר שכבה attention מקודד-מפענה (encoder-decoder) והוא חולפה בשכבה self-attention.

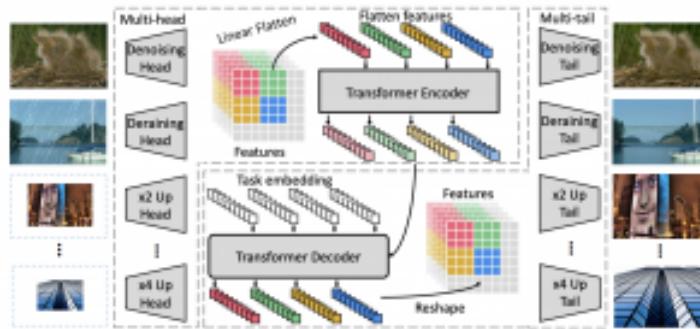


Figure 2. The diagram of the proposed image transformer (IPT). The IPT model consists of multi-head and multi-tail for different tasks and a shared transformer body including encoder and decoder. The input images are first converted to visual features and then divided into patches as visual words for subsequent processing. The resulting images with high visual quality are reconstructed by assembling output patches.

הערה על ארכיטקטורת המפענה: העדר שכבה attention מקודד-מפענה במפענה של IPT נראה לי קצת לא הגיוני. מבנה המשימות של עיבוד תמונה level-low (כמו ניקוי ראש או סופר-רזולציה) דומה לזה של התרגומים האוטומטי.cidoo הטרנספורמר הציג ביציעים טובים מאוד במשימות תרגום כאשר שכבת attention מקודד-מפענה משחקת תפקיד מאד חשוב בהבנת/כימות קשרים בין המשפט המקורי לתרגום שלו.

כמו שאמרנו בנוסף לאימון של IPT על מספר משימות level-low בו-זמןית, המאמר מציע לבצע אימון ניגודי' במטרה לשפר את הייצוג התמונה. אך בוואו קודם כל נרען מה זה שיטת אימון (למקרה ניגודית).

עקרונות הלמידה הניגודית: העקרון החשוב של גישה זו מניח שייצוגים של דוגמאות דומות צריכים להיות קרובים, כאשר שייצוגים של דוגמאות לא דומות צריכים להיות רחוקים. פונקציית המטרה בלימדה הניגודית מנסה למקסם את היחס בין אקספוננט של דמיון של זוג דוגמאות קרובות לסכום הדמיונות בין שני דוגמאות רנדומליות (זוגות שליליים)

המאמר משתמש בגרסה הסטנדרטיבית של הלוס הניגודי', אשר הדמיון בין ייצוגים מוגדר כדמיון קוסינוס (cosine similarity). דוגמאות קרובות כאן זה למעשה פאטצ'ים של אותה תמונה, אשר כל זוג של פאטצ'ים מת湊נות שונות מוגדר כשלילי'.

איך מאמנים IPT:

- **אימון מקודדים:** לוקחים את הדאטהסט של ImageNet ויצרים ממנו דוגמאות למשימות עליהן מאומן IPT. למשל למשימה של ניקוי ראש רגיל הם יוצרים תמונות מורעשות עי"ה הוספה של ראש לבן לתמונה כאשר המשימה היא לשחזר את התמונה המקורי.

- מאמנים את IPT למשימות low-level המשמשות כהמשך כל באטץ' מכל דוגמאות למשימה אחת בלבד (בשביל לה חסוך זמן חישוב עקב שימוש בשכבה ראש ו בשכבה נוספת בלבד).
- מבצעים במידה ניגודית על Imagenet. המאמר לא מצין מה סדר בין אימון למשימות low-level ו בין האימון הניגודי. אני משער כי הם מtbody'ם בלבד (גגיד באטץ' אחד עבור משימת low-level ובאטץ' אחד עבור הלמידה הניגודית).
- יכול של IPT מtbody'ם על DATAHESST מטרה.

לבסוף נציג כי פונקציית LOSS למשימות low-level הינה 1.1.

הישגי מאמר:

המאמר מצליך להראות שיפור בBITSUPIM עבור מספר משימות עיבוד תמונה low-level כמו סופר-רזולוציה, ניקוי רעש לבן והסרת רעש גשם (deraining) עבור כמה DATAHESSTים. עבור כל משימה לוקחים IPT מאומן ומכללים אותו על DATAHESST נתון.

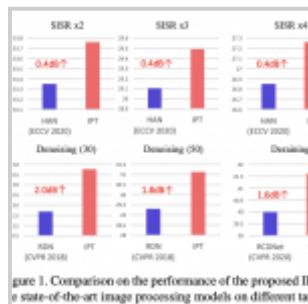


figure 1. Comparison on the performance of the proposed IP and state-of-the-art image processing models on different tasks

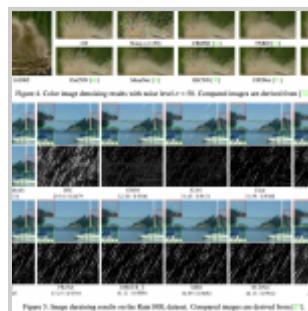


Figure 2. Color image deraining results with noise level = 30. Compares images are deraining [13]

Figure 3. Stego deraining results on the Rain-HSI dataset. Compares images are derained [13]

DATAHESSTIM:

Set5 , Set14, B100, Urban100 ,DIV2K

נ.ב.

מאמר מציע שיטה לאימון מקדים של רשת נוירונים עבור משימות low-level מרובות. הארכיטקטורה שלהם כוללת מוקודד ומפנעעה של הטרנספורמר הסטנדרטי ורשתות המפיקות פיצרים ייעודיים לכל משימה. המאמר מראה שיפור על מספר רב של שיטות SOTA והגישה גראית די מבטיחה. הייתה רצוחה להוכיח הוכחת עליליות

טיפה נוספת מבוססת על נתונים מדומינניים מגוונים יותר. גם הנקוד לא שותף לזה תמיד מאכזב. בקיצור מאמר מעניין אך נראה קצת לא מבושל למורח שהרעיון שהוא מציע נראה די חדשני ומעניין.

Review 34: Identifying Mislabeled Data using the Area Under the Margin Ranking

פינט הסוקר:

המלצת קריאה ממיק: כמעט חובה – (לא חובה אבל קרובה לזה 😊).

בהירות כתיבה: גבוהה

רמת היכרות עם כלים מתמטיים וטכניקות של DL/ML הנדרשים להבנת מאמר: היכרות בסיסית עם מושגים יסוד של הלמידה העמוקה (בעיקר אלו הקשורות לאימון של רשתות נירונים).

ישומים פרקטיים אפשריים: אופטימיזציה של תהליכי אימון של רשתות נירונים עי"ז זיהוי של דוגמאות מתיוגות תוך כדי האימון.

פרטי מאמר:

لينק למאמר: [זמן להורדה](#).

لينك לקוד: [כאן](#).

פורסם בתאריך: 23.12.2021, בארכיב.

הוזג בכנו: NeurIPS 2020.

תחומי מאמר:

- זיהוי דוגמאות בעלות לייבלים שגויים בתהליכי אימון של רשתות נירונים.

כלים מתמטיים הסימוניים:

- לוגיטים (logits): פלט של השכבה האחורונה של רשת סיווג (לפני הנרמול sigmoid).

תחומיים בהם ניתן להשתמש בגישה המוצעת:

- למידה semi-supervised.

תמצית מאמר:

אחד הגורמים המרכזיים שימושיים על ביצועים של רשות ניירונים הינו יכולות של הדטה סט שעליו הרשות מאומנת. חלק לא מבוטל של מדטה סטים הליבלים הינם "חלשים" (לא מדויקים) כי התיאוג בוצע דרך שימוש במסתני פרוקסוי או דרך דפי האינטרנט. זיהוי דוגמאות עם לייבלים מוטעים עשוי לשפר את יכולת הכללה של רשות וגם תוריד את רמת הזיכרון שלה (memorization).

מכיוון שרשות ניירונים העשויות הין בעלות מספר גובה של פרמטרים, נדרש דטה סטים גדולים מאוד בשבייל לאמן אותן. עבור רוב הדטה סטים לא ניתן (או מאוד יקר) לעבור עליהם במטרה להזיהות דוגמאות המתויגות בצורה שגוייה. עקב לכך יש צורך בפיתוח גישות אוטומטיות (ללא התרבות בני אדם) לזיהוי של דוגמאות כאלה.

המאמר מציע שיטה לזיהוי אוטומטי (ללא התרבות אנושית) של דוגמאות עם לייבלים שגויים במהלך האימון של רשות ניירונים. השיטה מנצלת את המידע על לוגיטים (logits) של דוגמאות לאורך אימון הרשות לזיהוי של דוגמאות עם לייבלים מוטעים. המטריקה שהם משתמשים בה נקראת שטח מתחת לשוליות (area under the curve – AUC). יותר קונקרטית, לדוגמא נתונה AUM מודד את ההפרש הממוצע על פני כל האפוקים בתהילן אימון הרשות, בין ערכי הלוגיט של הקטגוריה המתאימה לייבל שאיתו הדוגמא מתויגת, לבין המקסימום של כל ערכי הלוגיטים של הקטגוריות האחרות.



Figure 1: Images from MNIST (left) and ImageNet (right) with lowest Area Under the Margin (AUM) ranking (most likely to be mislabeled). AUMs are computed with LeNet/ResNet-30 models.

פינת האינטואיציה: עבור דוגמאות מתויגות נכון הפרשים אלו אמורים לאורך האימון כי יכולת הכללה (הנובנית על סמך דוגמאות עם אותו לייבל) של הרשות עולה ככל שהיא אימונם מתקדם. לעומת זאת בדוגמאות המתויגות בצורה שגוייה הרשות לא הצליחה לנצל את העלייה ביכולת הכללה שלה ו- AUM לא "מתורגם" ככל שהיא אימונם מתקדם. הסיבה לכך טמונה בעובדה שהרשות "רוואה" שדוגמא מתויגת עם ליibel `wrong`, דומה לדוגמאות הנושאות ליibel אחר (הנקון) `cor`_`cor` ומנסה לדוחוף את הלוגיט המתאים לו- `cor`_`cor` לעללה שהגורם לירידה במריגין של הדוגמא. אך באופן אינטואיטיבי AUM עבור דוגמאות "נכונות" אמר להיות יותר גבוהה מזה של הדוגמאות "הלא נכונות".

המאמר מציע לנצל את האינטואיציה זו ולזיהות דוגמאות שגויות על בסיס ה- AUM שלהם. מכיוון שאנחנו לא יודעים מה האחיזה של דוגמאות "לא נכוןות" בדטה סט נשאלת השאלה: איך נבחר את ערך הסף של AUM המבדיל בין דוגמאות נכוןות ללא נכוןות. המאמר מציע לבנות קטגוריה מלאכותית מתוך הדטה סט שעיל בסיס ה- AUM שלה הסף הזה נבחר.

הסבר של רעיונות בסיסיים:

כמו שכבר אמרנו עבור דוגמא x ואפוק \hat{z} , השול (margin) מחושב כהפרש בין הלוגיט של הליבל שאיתו הדוגמא מתואגת לבין המקסימום בין הלוגיטים של כל הליבלים האחרים. AUM עבור דוגמא x מוגדר כממוצע של הפרשיהם אלו על פני כל האפוקים. אתם תשאלו – מה הקשר של הממוצע זהה לשטח מתחת לגרף של מריג'נים: כדי להבין זאת מספיק להביט באיזור המצויר: שטח מתחת לשול משוערך ע"י הממוצע של השולים על פני האפוקים של אימון (אלו שעדיין לא הספיקו לשוכח חדו"א 1 יכולם לראות שהממוצע זה הינו סכום דרכו של הפונקציה מוגדרת על האפוקים והערכים שלה הם המרג'ינים). קל לראות שם המרג'ין של דוגמא הינו מספר חיובי גבוה אז הרשות מצליחה לחזות נכון את הליבל של דוגמא זאת. לעומת זאת השול שליל גבוה מכך שהרשות רואה את הדוגמא כדומה לדוגמאות המתויגות עם ליבל אחר (החיזוי שלה יהיה כנובן לא נכון עבור דוגמא זו).

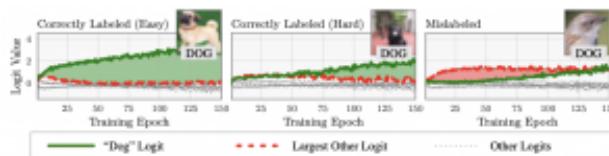


Figure 2: Illustration of the *Area Under the Margin* (AUM) metric. The graphs display logit trajectories for easy-to-learn dogs (left), hard-to-learn dogs (middle), and birds mislabeled as DOGS (right). (Each plot's logits are averaged from 50 CIFAR10 training samples, 40% label noise.) AUM is the shaded region between the dog logit and the largest other logit. Green/red regions represent positive/negative AUMs. Correctly-labeled samples have larger AUMs than mislabeled samples.

הערה: מספר עבודות הוכיחו שגדיל ממוצע של מרג'ין יכולת הכללה של הרשות: ככלומר ככל שהשול הממוצע גבוה יותר, יכולת הכללה של הרשות עשויה להיות טסקירה זו היא חלק מפינה קבועה בה אני סוקר מאמרם חשוביים בתחום ה-ML/DL, וכותב גרסה פשוטה וברורה יותר שלהם בעברית. במידה ותרצו לקרוא את המאמרים הנוספים שסיכמתי, אתם מוזמנים לבדוק את העמוד שマーיך אותם תחת השם [deepnightlearners](#).

Review 35: Regularizing Towards Permutation Invariance in Recurrent Models

פינת הסוקר:

המלצת קריאה ממייק: כמעט חובה (לא חייבים אף ממש מומלץ).

בהירות כתיבה: גבוהה.

רמת היכרות עם כלים מתמטיים וטכניקות של ML/DL הנדרשים להבנת מאמר: בניית מינוס – צריך להבין מה זה RNN ותכונותיו הבסיסיות. בנוסף מומלץ לຽunk את הידע הבסיסי בקומבינטוריקה (תמורות) ובתורת הקבוצות (מושגי יסוד).

ישומים פרקטיים אפשריים: ניתן להשתמש בטכנית זו בשבייל מושימות עיבוד סדרות אינוריאנטיות (באופן מלא או חלק) לסדר איבריהן כמו מושימות זיהוי של ענני נקודות, מציאת דמיון בין סטים של אובייקטים, זיהוי אוטומטי ECC וכדומה.

פרטי מאמר:

לינק למאמר: [זמן להורדה](#).

לינק לקוד: לא הצלחתי לאתר.

פורסם בתאריך: 25.12.20, בארכ'יב.

הציג בכנס: NeurIPS 2020.

תחומי מאמר:

- רשותות מסווג RNN.
- מושימות אינוריאנטיות לסדר של קלט.

כלים מתמטיים, מושגים וסימונים:

- תמורה (פרמוטציה) של סדרת קלט (יסומן c-d).

תמצית מאמר:

קיימות לא מעט בעיות בתחום למידת מכונה שהן אינוריאנטיות לסדר של הקלט, כולל לא תלויות באיזה סדר אנו מכניםים את הקלט לרשות. בין מושימות כאלו גמוניות מושימות כמו זיהוי סגמנטציה בענני נקודות במידול 3D, מציאת דמיון בין סטים של אובייקטים (למשל מציאה נוספת תמונה הכי דומות) ועוד. קיימות מספר גישות לביעות מהסוג הזה בلمידה עמוקה. אחת הגישות היא בניית רשת נירונית שהיא אינוריאנטית לסדר של קלט באופן אינהרנטי. אחת הדוגמאות לארכיטקטורות כאלו הינה "סט טרנספורמר" שבilibio נמצא מגנון של "self-attention" המוכר לנו מהתרנספורמר הקלסטי המשמש כמעט בכל תחומי ברירת המחדל לשימושות NLP". שימושו לבתרנספורמר המקורי אינו (!! אינוריאנטי לתמורות של הקלט כי הוא מכיל רכיב קבוע המיקום positional encoding). בנוסף הארכיטקטורה של הטרנספורמר מכילה כמה שכבות שלא מקיימות את תנאי האינוריאנטיות לתמורות כמו שכבת FC. רשותות נירונית בסגן RNN (והשכלולים שלו כמו LSTM, GRU וconditional) המיועדת לעיבוד דאטא סדרתי כמו שפות טבעיות, אותות דיבור ורטוי וידאו (החלק הטemporלי) מבוסנת אין אינוריאנטיות לסדר של הקלט מהסיבה פשוטה שיש להם "סדר טבעי אינהרנטי".

از נשאלת השאלה איך נוכל לבנות ארכיטקטורת רשת שהיא אינויריאנטית לtransforms ולא בעלת סיבוכיות חישובית גבוהה (ריבועית ביחס לאורך הקלט) כמו הטרנספוררים. זה תחום מחקר פעיל שהספיק להניב כמה תוצאות מעניינות. למשל מחברי DeepSets חקרו תכונות של פונקציות אינויריאנטיות לtransforms והוכיחו שככל פונקציה f צזו ניתן לתאר (למודל) ע"י שתי רשותות $R1$ ו- $R2$ באופן הבא:

- עבור כל איבר בסדרת קלט (x_1, \dots, x_n) = x מחשבים את הפלט של $(R1)_i(x)$.
- סוכנים את כל הערכים של $(R2)_i(x)$ מהשלב הקודם
- מעבירים את הסכום הזה דרך רשות $R2$

דרך אגב, המאמר הנזכר מוכיח משפט מאוד מעניין הטוען שככל אורך סדרה K גדול מ- 4, קיימות פונקציה אינויריאנטית לtransforms כאשר ניתן למשם אותה עם 3 נוירונים מושתרים בלבד כאשר DeepSets צריך לפחות K נוירונים בשבייל למשם אותה. לעומת זאת, המחברים רמזים שהגישה של DeepSets עלולה להיות לא מאוד יעילה עבור שימושות מסוימות (נכון שהמשפט בונה רק שימושה אחת צזו אך לדעתי יש משפה ורבה של שימושות כאלה).

בעבודות אחרות הציעו למצע פלטים של הרשת עבור כל הפרמוטציות של הפלט. מבון רשות צזו הינה אינויריאנטית לtransforms אך אינה יסימה בסקלר (עבור סדרות ארוכות) עקב סיבוכיותה המעריצית. כמו שכבר הזכירתי ארכיטקטורת self-attention שזיהה אינויריאנטיות לtransforms אולם גם היא בעלת סיבוכיות ריבועית ביחס לאורך הקלט שגם מקשה על יישומה למשימות עם קלט ארוך.

המאמר מציע גישה אחרת שאומרת כך: בואו ניקח רשות שהיא לא אינויריאנטית לפרמוטציות ונאלץ אותה להיות צזו בעזרת רגולרייזציה. הר' אם ניקח רשות RNN ו"נאלי" אותה כך להוציא אותה פלט צזו הינה כל פרמוטציה אפשרית של כל סדרות הקלט מהדעתה, אך האינטואיציה אומרת לנו אמורים לקבל רשות "קרובה לאינויריאנטיות לtransforms של הקלט". אבל צריך לזכור שדריך אימון צזו לא מבטיחה אינויריאנטיות לtransforms(!!) לא משנה עד כמה גדול סט האימון (פרט למקרה הלא מעניין שט האימון מכיל את כל הסדרות האפשריות עבור משימה זו). הסיבה היא, שלא ברור עד כמה אופן אימון צזו יודע "להכליל". לעומת זאת כמה אינויריאנטיות לפרמוטציות על הדוגמאות מסט האימון מועברת גם לטסט סט. מדובר באמת בשאלת מודול טריוויאלית.

המאמר מציע לקחת את הגישה ההז אובל במקום "לאלי" RNN להיות אינויריאנטית על כל הפרמוטציות של סדרת קלטים, המאמר מציע "לאלי" אותו להיות אינויריאנטי על תת-קובוצה z_{pr}^P של הפרמוטציות. תת קבוצה זו מורכבת מכל התמורות ששווות לפרמוטציה זהה פרט לשני מקומות. לעומת כל תמורה מ- z_{pr}^P היא למעשה חלקו של שני איברים בסדרת קלט המקורי. קל לראות שגם הרשות אינויריאנטית על כל פרמוטציה k מ- z_{pr}^P לכל הקלטים אז היא אינויריאנטית לכל התמורות האפשריות של הקלט. השיטה המוצעת קיבלה – SIRE – Subset Invariant REgularizer.

הערה: הסיבה שהמחברים בחר בארכיטקטורה של NNN נובעת מהמבנה הייחודי שלו, כאשר המצב המוסתר ($t+1$) s מוגדר כפונקציה של המצב t s בזמן t והקלט x_t . זה מבנה מואוד נוח לביעות אינויריאנטיות לtransforms (כמו מקסימום של סדרה) כי המצב t s יכול לצבור את הסטטיסטיקה על הקלט עד זמן t בזרה אינויריאנטית.

הסבר של רעיונות בסיסיים:

בתכלס המאמר מציע להוסיף לLOS הרגיל של המשימה איבר רגולרייזציה השווה להפרש הריבועי בין פלט הרשות, עבור סדרה המקורית, לפט הרשות עבור הקלט אחריו פרמוטציה m_{pr}^P . פרמוטציה זו נבחרת באקראי מכל

הפרמטריזציות האפשריות מ- $z_{\text{C}}P$. המאמר לא מפרט האם מוסיפים הפרש צזה עבור תמורה אחת בלבד או בוחרים כמה כאלו באקרהי (אני חשב שמספר התמורות באיבר רגולרייזציה צריך להיות אדפטיבי להיקב ע"י היחס בין הלוס על המשימה וגודל הפרש ממוצע על כמהTamorphot מ- $z_{\text{C}}P$).

הישגי מאמר:

המאמר מראה ש- SIRE מפגין ביצועים יותר טובים מ- DeepSets במספר שימושות (הם לא השוו את הביצועים על מול Set Transformer מבוסס על מנגןון *attention* כנראה בגל הסיבוכיות הריבועית שלו).

1. חישוב של parity של סדרה (סכום הזוגות).
2. חישוב של סכום, טווח (הפרש בין האיבר המקסימלי למינימלי) שונות של סדרה.
3. דיזי אובייקטים וסיווג בענייני נקודות.
4. חצי טווח של סדרה (הפרש של המקסימום של החצי הראשון של סדרה והמינימום של החצי השני שלה). משימה זו סמי-אינוריאנטית לתמורות (עבור סט מאד גדול של תמורות אך לא עבר כלון, התוצאה לא משתנה).
5. שימוש סיווג על Perturbed MNIST, הנווצר מ- Locally Perturbed MNIST רגיל ע"י החלוף של פיקסלים שכנים רנדומליים. משימה זו כמובן סמי-אינוריאנטית לתמורות.

בכל המשימות הם הראו יתרון של SIRE על DeepSets פרט למשימה אחת עבור ענייני נקודות. מעניין שעבור Locally Disturbed MNIST הם ניטו RNN רגיל שקיבל דיוק של 87% בזמן של CNN סטנדרטי הפגין דיוק של 96%. אולם, אחרי ההוספה של איבר הרגולרייזציה שלהם (כנראה כן הם לא לקחו כל פרמטריזציה אלא רק כאלה שמתאימות לאופן ייצור הדטה סט) הדיוק הגיע ל- 97.7%.

ג.ב.

מאמר מציע שיטה להתאים את RNN למשימות אינוריאנטיות לתמורות של קלט. השיטה מאוד אינטואיטיבית, קלה להבנה ומוסברת היטב – כיף לסקור צזה. אבל לי עלו כמה שאלות שלא מצאת תשובה עליהם במאמר. למשל לא ברור לי כמה Tamorphot מ- $z_{\text{C}}P$ אני צריך בשביל לאמן סדרות עד גודל מסוים (נגיד אני רוצה לאמן רשת לחישוב טווח של סדרה עד גודל 100 – כמה אפוקים אני צריך בשביל להתכנס לתוצאה טובה). בנוסף גם מעניין לראות אנהיזה של השפעת אורכי הסדרות עליהם מאמנים RNN עם SIRE על הביצועים). המאמר גם הציג השוואה של השיטה שלהם מול גישה אחת בלבד DeepSets ובוקר על בעיות עצום – הייתה רוצה לראות איך הביצועים של SIRE מול שיטות כמו set transformer ודומה.

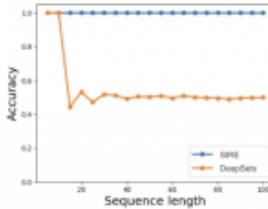


Figure 1: Test accuracy as a function of sequence length for learning parity, using DeepSets and RNNs.

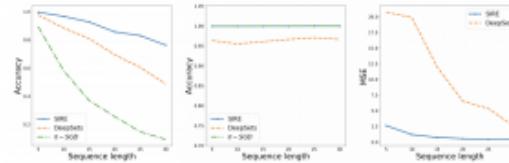


Figure 2: Test prediction accuracy (zero-one error) of saw (left) and range (center). For the variance experiment we report mean square error (as in Murphy et al. [2018]).

Method	100 pts	1000 pts	5000 pts
DeepSets	0.825	0.872	0.90
SIRE	0.835	0.878	0.899

Table 1: Point cloud classification results.

Review 36: Sequence-to-Sequence Contrastive Learning for Text Recognition

פינט הסוקר:

המלצת קרייה ממייק: כמעט חובה (לא חייבים אך מומלץ בחום לחסידי למידת הייצוג ואובי OCR).

בahirot כתיבה: גבוהה.

רמת היכרות עם כלים מתמטיים וטכניקות של DL/ML הנדרשים להבנת מאמר: בינוי (נדרשת הבנה מסוימת במושגי למידת הייצוג).

ישומים פרקטיים אפשריים: שיפור ביצועים למשימות OCR כמו זיהוי לוחות רישוי, זיהוי של תמרורים עבור מערכות רכב אוטונומי, הקטנת גודל סט אימון מתייג הנדרש לרמת ביצועים נתונה.

פרטי מאמר:

lienק למאמר: [זמן להורדה](#).

לינק לקובץ: לא הצלחתי לאתר.

פורסם בתאריך: 20.12.20, בארכיב.

הוגג בכנסו: NeurIPS 2020

תחומי מאמר:

- למידת ייצוגים במימד נמוך למשימות זיהוי טקסט (כתב יד) בתמונה.
- למידה ניגודית (CL - contrastive learning) למשימות מיפוי סדרה לסדרה (sequence-to-sequence tasks - StST) .

כליים מתמטיים, טכניקות, מושגים וסימונים:

- LOSS ניגודי (contrastive loss).
- אוגמנטציה של דאטא לצירעה של דוגמאות "דומות".
- רשתות לעיבוד שדאטה סדרתי (sequential) כמו LSTM.

מבוא והסבר כללי על תחום המאמר:

שיטות לבניית ייצוג במימד נמוך של דאטא עבור דאטאטיטים לא מותגים, המבוססות על למידה ניגודית, תפזו פופולריות רבה בשנים האחרונות. לאחרונה שיטות אלו הצליחו לבנות ייצוגים מאוד חזקים שלא נופלים באיכות מלאו שנבנו במהלך אימון של מודלים עמוקים עם דאטאטיטים מותגים. יתרה מזאת, ישום גישה זו בשילוב של דאטאטיט מותיג לא גדול (למידה semi-supervised) הצליח להזניק את הביצועים במגן של משימות כמו סיווג, סגמנטציה, זיהוי אובייקטים ואחרות מעבר לאלו של המודלים שנבנו עבור דאטאטיטים מותגים גדולים.

הרבית שיטות אלו הוצעו במקור עבור דומין של התמונות (SimCLR, BYoL, MoCo ועוד). נציין שימושם לאחרונה פורסמו עבודות המציאות גישות מבוססות על הלמידה הניגודית גם בדומיניים אחרים כ**למידה ניגודית לסדרות זמן** ו- **למידה ניגודית לדומין הוידאו**. לעומת זאת רוב השיטות העכשוויות לזרוי כתוב יד הין שיטות למידה supervised בסיסיות. הסיבה לכך טמונה בעובדה שגישות למידה ניגודית מצטיינות ביצירת ייצוגים חזקים לשימות כמו סיווג או זיהוי שביהם כל דוגמא הינה "פיסה אוטומטית" (במילים אחרות הדוגמא "מהו זה מקרה אחד"). במקרה כזה דוגמאות "חיבויות" (קרובות) לדוגמא נתונה **X** נוצרות עי"י הפעלת אוגמנטציות שונות על **X**, כאשר כל שאר הדוגמאות הין "שליליות" ל **X** (רחוקות) באופן אינגרנטי. המצב במשימות זיהוי של כתוב יד שונה כי תרлик החיזוי של מילה כתובה הינו סדרתי באופן טבעי (כמו כל זיהוי של טקסט). כל מילה כתובה מורכבת מאותיות ונראה שהכי הגיוני למדל אותה כסדרה של פאצ'ים (נקרא לזה גם פרימרים) סמכים של טמונה שככל אחד מכם מייצג תת-מילה/אות חלק מאות. עקב כך הצורה הסטנדרטיבית של הלמידה הניגודית אינה ישימה עבור תമונות המכילות כתוב יד (למעשה אוגמנטציה של טמונה עם מילה כתובה **X** עלול להפוך אותה ל'לא קרובה' ל- **X** לאחר והיא משנה את ה"סדר הטבעי" של האותיות).

הסביר על מושגים חשובים במאמר:

עקרונות הלמידה הניגודית: גישה זו מסתמכת על ההנחה שיצוגים של דוגמאות קרובות צריכים להיות קרובים, בזמן שיצוגים של דוגמאות לא קשורות (נקראות שליליות) צריכים להיות רחוקים. בשביל לבנות פונקציית מטרת לשיטת למידה ניגודית לוגדים זוג של דוגמאות קרובות (למשל שתי אוגמנטיות של הזוג דוגמא) ומספר דוגמאות רנדומליות ומנסים למקסם את היחס בין אקספונט של דמיון של הזוג הקרוב לסכום הדמיונות בין שני דוגמאות רנדומליות.

תמצית מאמרה:

המאמר מציע גישה של למידה ניגודית, הנקראת SeqCLR (הगרסה הסדרתית של SimCLR) המותאמת למשימות בעלות אופי סדרתי כגון זיהוי של טקסט בתמונה. בגדי הרעיון הוא ליצור מתמונה **סדרה(!!!)** של אובייקטים כאשר כל אובייקט מורכב מספר פאצ'ים סמכים ולמוד יציג כל אובייקט כזה באמצעות למידה ניגודית. נגשים שהסדר בין האובייקטים נשמר וזה הסיבה שהשתמשתי במונח "סדרה"(!!!) ולא במונח "סדר"(!!!) בתיאור. כמובן ניתן ליצור מספר שונה של אובייקטים לכל תמונה. אז איך בעצם עבדת הלמידה הניגודית כאן, ככלمر אף בונים זוגות "חיוביים" וזוגות "שליליים" של דוגמאות(זהה הבסיס של כל גישת הלמידה הניגודית). אז התהילה נראה כך.

- מפעלים שתי אוגמנטיות על תמונה **x** ויצרים תמונות **1_x** ו- **2_x**. שימוש לב שלא כל שיטת אוגמנטיות משתמשים בה בלמידה הניגודית הסטנדרטיבית הדומין התמונות, מתאימה כאן. למשל אוגמנטיות מסווג הזזה אופקית והיפוך אינם מתאימים כאן כי הן עלולות "לשנות את הסדר בין פרימיום (הסיבה לכך מפורטת באחד הסעיפים הבאים).
- מחלקים את **1_x** ו- **2_x** לסדרות של אובייקטים **{i_2_x}** ו- **{i_1_x}** (הסדרות הן באותו אורך). נזכיר כל אובייקט מורכב מפרימיוםים (פאצ'ים) סמכים.
- בונים ייצוגים מ **{i_2_z}** ו- **{i_1_z}** לכל האובייקטים שנבנו ע"י העברתם דרך כמה רשותות נירונים עוקבות (היצוגים נבנים בצורה זהה עבור **1_x** ו- **2_x**).
- כל זוג חיבי בניו מהאובייקט נבנה מייצוג **k_1_z** ו- **k_2_z** עבור k-ים שונים כאשר ייצוגים **i_1_z** ו- **j_2_z** עם אינדקסים שונים ויצוגים של האובייקטים של הדוגמאות האחרונות מהבאטץ', מהווים זוגות של דוגמאות שליליות. שימוש לב שכך שמספר האובייקטים, נוצרם מתמונה בודדת, גובה יותר, צרי באטץ' יותר קטן בש سبيل ליצור אותה כמות של זוגות של דוגמאות שליליות לדוגמא **x**. בנוסף יש לי תהושה שהדוגמאות השליליות הנוצרות מאותה תמונה "מלאכות" את המודל ללמידה בין אובייקטים שונים בתוכן (אותיות שונות) אך דומים מבחינת הסגנון שתורם חיובית לעוצמת הייצוג. בנייה זו ממחישה למה ניתן להשתמש רק באוגמנטיות שלא משנה את הסדר של האובייקטים(פרימיוםים) בתמונה.

הסבר של רשותות בסיסיים:

אחר שבניים את עיקרי SeqCLR, נותר רק לפרט איך נבנים הייצוגים **{i_z}** מסדרת הפרימיוםים של תמונה **x**. בנייה זו נעשית דרך שימוש בכמה רשותות נירונים עוקבות שמפורטות בסעיף הבא:
טהיליך בניית ייצוגים עבור תמונה x:

- **בנייה ייצוגים התחלתיים של פרימיום:** קודם כל מחלקים תמונה ל- D פרימיום. אחר כך יש שתי אופציות: הראשונה היא להעביר כל פרימום דרך רשות ייצוג סטנדרטיבית לתמונות ולהשתמש בתוצאה בתור ייצוג של כל פרימום. האפשרות השנייה היא לבנות ייצוגים המנצלים את הקשרים בין הייצוגים של הפרימומים השונים מהשלב הראשון (ייצוג קוונטקטואלי או הקשי'). המאמר

מציע לעשות זאת עם LSTM זו כיוון (כאן ברוח הזמן היתי מציע להשתמש באיזו גרסה "קלה חישובית" של הטרנספורמר). ניתן כמובן לשלב את שני היצוגים האלה כקלט לשלב הבא.

- **בנייה ייצוגים של אובייקטים:** המטרה בשלב זה הינה ליצור קלט לחישוב של הלס הניגודי. המאמר מציע 3 דרכים לבנות אותם מהיצוגים של הפריים שהתקבלו מהשלב הראשון:
 1. למצוות כל ייצוגים ולקיים וקטור יציג אחד \mathbf{z} עבור התמונה ולקבל אובייקט אחד לכל תמונה. במקרה זה זוגות שליליים נבנים רק בין דוגמאות שונות בבאץ'.
 2. לבנות \mathbf{c}_T אובייקטים מ- T פריים כאשר $\mathbf{c}_T > T$ (פולינג) המאפשר בניית זוגות דוגמאות שליליות גם מאובייקטים אותן אותה תמונה.
 3. כל פריים הופך לאובייקט.

- **חישוב של לוס ניגודי** מהיצוגים שהתקבלו בסעיף הקודם.

הישג מאמר:

המאמר הצלח להוכיח שהיצוגים שנבנו באמצעות SeqCLR מסוגלים להביא לשיפור ביצועים במקרים מסוימים על דאטאסתים. הבדיקות נעשו בצורה סטנדרטיבית עבור שיטות מסווג זהה: מקפאים את משקל המודל (שבונה את יציג הדטה), מוסיפים שכבת לינארית למודל ומאמנים אותה על דטה מתוג. SeqCLR הגיע לביצועים יותר טובים גם עבור משימות semi-supervised (כלומר האימון של השכבה הלינארית מתבצע רק על אוחז מסוים של דוגמאות מדאטאסט מותוג).

דאטאסטים:

- כתוב יד באנגלית: IAM, CVL .
- כתוב יד בצרפתית S.RIMES .
- זיהוי טקסט בתמונה: SyntText, IIT5K, ICO3 .

ג.ב.

המאמר מכיל את שיטת הלמידה הניגודית לדיאטה בעלת אופי סדרתי כמו טקסט בתמונה. השיטה מאוד אינטואטיבית וקל להשתכנע שהיא אכן מגיעה לתוצאות טובות. השיטה נבדקה על לא מעט דאטאסטים בשני דומיננסים שונים (כתב יד וtekst בתמונה - OCR) ונמצאה עדיפה על מתחרה. החלק היחיד ששוכר לי בשבי להשתכנע זה הקוד שלא נמצא באקריב - בתקווה שיוסיף בקרוב.

Review 37: Teaching with Commentaries

פינת הסוקר:

המלצת קריאה ממיק: מומלץ לאוהבי מטה-למידה ובורי רקע בחזו"א 2 מתקדם.

bahiorot ctiyeh: בינוית.

רמת היכרות עם כלים מתמטיים וטכניקות של ML/DL הנדרשים להבנת מאמר: רקע טוב בתחום מטה-למידה, חזו"א ברמה גבוהה.

ישומים פרקטיים אפשריים: ניתן להשתמש בגישה זו למשל לזרחי דוגמאות המשפיעות ביותר על הביצועים או איתור פאצ'ים בתמונות מהדאטasset החשובים למשימה במהלך האימון של הרשת.

פרטי מאמר:

[لينק למאמר: זמין להורדה.](#)

[لينك לקוד:](#) לא הצלחתי לאתר.

[פורסם בתאריך:](#) 5.11.20, בארכיב.

[ייצוג בכנס:](#) ICLR 2021.

תחומי מאמר:

- שיטות אימון של רשתות נוירונים.
- שיטות מטה למידה (meta-learning) בתחום רשתות הנוירונים.

כליים מתמטיים, מושגים וסימונים:

- משפט הפונקציה הסתומה.
- חישוב נגזרת של פונקציה וקטורית דרך ההפכית של מטריצת הסיאן (hessian).
- קירוב ניומן (neumann) לחישוב ההפכית של אופרטור (מטריצה) לינארי.
- רשת לומדת פנימית (inner student network).
- רשת מלמדת (נקראת הרשת המפרשנת במאמר - commentary network).
- אימון פנימי/חיצוני (inner/outer optimization).
- מטה-אימון, (meta-training).

תמצית מאמר:

כמו שאותם בטח יודיעים, למורות הפעילות המחקרית האינטנסיבית בתחום הלמידה העמוקה עדין קיימות לא מעט סוגיות פתוחות בנושאי אימון, רגוליזציה והבנה של מה שקרה בתוך רשתות נוירונים עמוקות. השאלות הללו נוגעות בסוגיות בסיסיות כמו: איך לאמן רשתות לצורה יותר מהירה, איך להקטין את כמות הדadata הנדרשת לאימון, איך לשפר את יכולת הכללה וובسطיות של הרשתות.

אחד הגישות המעניינות שהוצעה לאחרונה שמנסה לתת מענה לשאלות אלו נקראת "לומדים ללמד" (learning to teach) המציע לבנות מנגנון חיצוני (רשת בדרך כלל) בשבייל לספק לרשת הלומדת תובנות (נקרא לזה גם מטה-ידע בהמשך) לגבי המשימה תוך כדי תהליך האימון. למשל מנגנון זהה יכול לבצע משקלול של דוגמאות במטרה לעזור לרשת הלומדת "לרכז את המאמץ" בדוגמאות החשובות. דוגמא אחרת של גישה זו יכולה להיות רשת עזר ה"מייצצת" איך לבנות דוגמאות (למשל ע"י ערבות של דוגמאות מדatta סט) הגורמים לרשת הלומדת לבנות יציג חזק של DATA.

מאמר זה מציע מסגרת כללית לגישה זו(הנקראת למידה עם פרשנויות) ומציע תהליך אחדי להפקה של מטה-מידע (פרשנות) ע"י רשות חיצונית (פרשנית) com_N , תוך כדי "הסקת מסקנות" העולות בתהליכי האימון של רשות st_N (נקראת הרשות הלומדת) על סט האימון. אך בואו נבון איך כל זה עובד בעצם? נניח שאנו רוצים למצוא איזושהי טרנספורמציה (לדוגמא משקל/ערובוב) של דוגמאות בדאטה סט במטרה לשפר את הביצועים של הרשות הלומדת st_N והמטרה של הרשות המפרשנת com_N הינה למצוא את הטרנספורמציה זו וזה למעשה מהו הפלט שלה Out_com . במקרה זה תהליך האימון מכיל את השלבים הבאים:

- אופטימיזציה פנימית: עבור סט משקלים נתון com_W של הרשות com_N , מאמנים את st_N (כמו איטרציות של GD על משקל GD_W). במקורה הזה מפעלים טרנספורמציה Out_com המופקת ע"י (N_com, W_com) על הדאטה של סט האימון ומאמנים את st_N עליו. הפלט של השלב הזה הוא המשקלים W_st_N .
- אופטימיזיה חיצונית: מחשבים את הלוס של st_N עם סט המשקלים W מהשלב הקודם על סט ולידציה שעובר טרנספורמציה הניתנת ע"י (N_com, W_com) . כאן מאמנים את N על משקל W_com כלומר מבצעים כמו איטרציות של GD אבל הפעם על משקל W_com .
- חוזרים על הצעדים אלו T פעמים כאשר T זה מספר האיטרציות של מטה אימון.

```

(1) Initialize commentary parameters  $\phi$  and student network parameters  $\theta$ 
(2) For  $M$  steps:
    (i) Compute the student network's training loss,  $\mathcal{L}_T(\theta, \phi)$ .
    (ii) Compute the gradient of this loss w.r.t the student parameters  $\theta$ .
    (iii) Perform a single gradient descent update on the parameters to obtain  $\hat{\theta}$  (Note this
          is implicitly a function of  $\phi$ , i.e.  $\hat{\theta}(\phi)$ ).
    (iv) Compute the student network's validation loss,  $\mathcal{L}_V(\phi)$ .
    (v) Compute  $\frac{\partial \mathcal{L}_V}{\partial \phi}$ .
    (vi) Approximately compute  $\frac{\partial \mathcal{L}_V}{\partial \phi}$  with equation 4, using a truncated Neumann series with a
          single term and implicit vector-Jacobian products [17].
    (vii) Compute the overall derivative  $\frac{\partial \mathcal{L}_V}{\partial \phi}$  using (v) and (vi), and update  $\phi$ .
    (viii) Set  $\theta \leftarrow \hat{\theta}$ .
(3) Output:  $\phi$ , the optimized parameters of the commentary.

```

הסבר של רעיונות בסיסיים:

קודם כל נזכיר כי הגרדיאנט של משקל com_N משערק את השינוי בלווי של st_N ביחס לשינוי במשקלים של com_N . אבל צריך לזכור שבשביל לחשב את הלוס של האופטימי של st_N עבור משקל com_N נתונים, המשקלים של st_N עוברים כמה איטרציות (אול' די הרבה) של GD במטרה למזרע את הלוס שלו. لكن כדי לחשב את הגרדיאנט של הלוס של st_N לפי משקל W_com צריך "לגלגל את כל האיטרציות על משקל st_N " - com_W שזה יכול להיות די כבד חישובית כאשר com_N הינה רשות גדולה.

גם אם נחליט להשתמש רק באיטרציה אחת בתהליכי האופטימיזציה הפנימית (זה מה שעשו במאמר) עדין של לנו בעיה עם חישוב הגרדיאנט של הלוס לפי com_W . הבעיה זו נובעת מההעבדה שగרדיאנט זה שווה למכפלה של הגרדיאנט של הלוס לפי st_W (שהזה ניתן לחשב אותו בצורה הסטנדרטית של גזירת הלוס של רשותות) והנגזרת של וקטור משקלים st_W לפי לוקטור משקלים com_W . נזכיר ש st_W תלוי ב- W_com בצורה לא מפורשת כי בשלב האופטימיזציה הפנימית st_W מחושב על הדאטה סט אחריו הפעלת עליון טרנספורמציה המוגדרת ע"י com_W . נגזרת זו היא בעצם מטריצה (W_com -ו- W_st הם וקטוריים) שמיימת עלולים להיות די גבוהים. נניח ש st_N ו- com_N הם רשותות לא גדולות בגודל של מיליון משקלים אז הנגזרת זו תהיה מטריצה בגודל מיליון על מיליון ותידרש כמות זיכרון עצומה בשבייל לאחסן אותה.

לכן המאמר מציע להשתמש במשפט הפונקציה הסתומה עבור הנגזרת של הלוֹס (של השלב החיצוני out_L) לפי W . משפט זה מאפשר לתאר את הנגזרת הבועיתית ע"י מכפלה של הופכית הסיאן של לואָס של השלב הראשון in_L לפי W_{st} והמטריצה של הנגזרות המעורבות לפי com_W ו- $-in_W$ של in_L . למי שלא זכר הסיאן זה מטריצה המורכבת מהנגזרות שנית in_L לפי הזוג של רכיבים של com_W ו- $-W_{st}$. צריך לזכור שהפירוק לעיל מתקיים בסביבת נקודה שבה הנגזרת של in_L לפי com_W מתאפסת. העובדה שלא ניתן למצאו אותה במדויק וזה יכול להשפיע בצורה שלילית על התהילה המתה-למידה.

גם אחרי הפירוק הזה יש לנו בעיה - והוא טמונה בחישוב של הופכית של הסיאן של in_L לפי W_{st} . אפילו עבור st_W בגודל יחסית לא גדול חשובו ההופכית (ולפעמים ההסיאן עצמו) יכול להיות מאוד ידרוש כמות עצומה של חיצון וזמן. בשайл להקל על היבט החישובי משתמשים בקירוב ניימן עבור ההופכית מוכפלת בגרדיאנט של in_L לפי W_{st} (מאמר של לוריין) תוך שימוש בצורת עדכון של GD. דרך אגב נוסחת ניימן מאתרת הופכית של אופרטור לינארי בתור טור אינסופי של החזקות שלה (מוזגות במינוס מטריצה היחידה). במאמר משתמשים רק באיבר אחד של קירוב זה.

הישגי מאמר:

במאמר מראים 3 דרכי להשתמש בגישה זו לשיפור תהיליך האימון של רשותות נירונים:

- חישוב של משקל על דוגמאות מسط האימון ע"י com_W (דוגמאות עם משקל גבוה משפיעים יותר על הלוֹס). מעניין שהם גם בדקו את הביצועים של השיטה שלהם בתרחיש למידת few-shot שזו שימושת מתה-למידה קלאסית. המטרה בלמידת few-shot היא לאמן רשות חיצונית (מתה) על מספר משימות (שנלמדות בפועל ע"י הרשות הפנימית) באמצעות למידה להפיק תכונות משותפות של כל המשימות (שימוש לביצה מקורה פרטיה של הפרדיגמה הכללית שהוצאה במאמר). כאשר מגיע משימה חדשה הרשות החיצונית מסוגלת לכידל את עצמה עם כמות קטנה של DATA המשמשה הזו. זו אחת דרכיהם לפטור בעיה זו והואאמנת רשות חיצונית לאתחול המשקלים של הרשות הפנימית שתאפשר לה להציג לביצועים טובים על משימה חדשה במספר איטרציות GD קטן. אז הם מראים שהשילוב של MAML עם משקל דוגמאות הבננה ע"י com_W גורם לשיפור ביצועים משמעותית. המאמר מראה שיפור ביצועים עבור הדטה סטים MinilmageNet ו- CUB200-2011.

- בניית של מקדים ערבות אופטימליים עבור הדוגמאות (בדומה ל mixup). כאן דוגמא מעורבתת לבנית סכום קמור של שתי דוגמאות: $2x_2 - (1 - x_1) + ax_1 = mix_x$ כאשר מטרת com_W הינה לחשב את מקדמי a האופטימליים לביצוע משימת סיווג(מאמר קשור המקורי מגרייל אותם מהתפלגות בטיה). מעניין שכן "למקרה עם פרשניות" מנצחת את $mixup$ ב- CIFAR10 ומציג ביצועים קצר פחות טובים ממנו על CIFAR10 (DATA סט יותר קטן).

- חישוב מסכות על תמונות לשיפור הפיצרים המופקים ע"י רשות. כאן com_W בעצם מחפשת אזורים "חשוביים" בתמונה שצדאי לרשות הלומדת להתרצע עליהם. המאמר מראה באופן ייזואלי שהמסכות שהוא מפיק אכן מתמקדות באזורי החשובים של תמונות ומראים באופן כמוות את עדיפותן על פני שיטות אחרות לבנית מסכות.

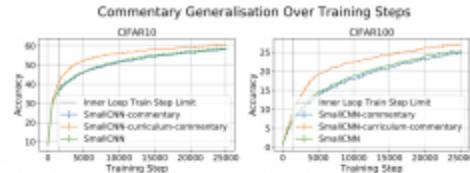


Figure 2: Example weight curricula can speed up training. Test-set accuracy curves on CIFAR10/100 when using curriculum commentaries, non-curriculum commentaries, and no commentaries, during student network training. The learned curriculum commentary network which generates pre-tutorial example weights results in learning speed improvements. This learning speed improvement holds when the student network is trained for many more steps than the number of inner loop update steps used during commentary network training (1000 steps). This demonstrates that the curriculums generalise to longer training times.

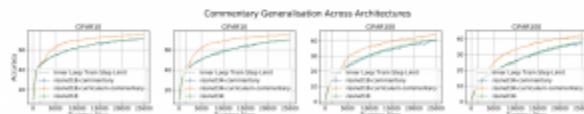


Figure 3: Example weight curricula generalise across network architectures. Using a learned curriculum commentary network trained with a simple 2-block CNN student, we apply these example weights to train two ResNet architectures. This gives improved test set accuracy curves for ResNet students also, indicating that the curriculums generalise across architecture.

ג.ב.

מאמר מציע מסגרת כללית לשיפור של תהליכי למידה של רשתות נוירונים ثنائية להשתמש בה למגוון רחב של משימות של הלמידה העומקה. הגישה שלהם גם עוזרת להפיק תובנות חדשות תוך כדי תהליכי האימון של רשתות. אני מני שעוד נשמע על שימושים רבים של גישה זו....

Review 38: Removing Bias in Multi-modal Classifiers: Regularization by Maximizing Functional Entropies

פינט הסוקר:

המלצת קריאה ממ"ק: מומלץ מאוד אך לא חובה (זהירות: מתמטיקה קצת קשוחה בפנים).

זהירות כתיבה: גבוהה.

רמת היכרות עם כלים מתמטיים וטכניקות של ML/DL הנדרשים להבנת מאמר: נדרש רקע מוצק בתורת האינפורמציה וכלים מאנגליזה פונקציונלית בנוסף להבנה عمוקה בסוגיות העולות באימון של מודלים מולטימודליים.

ישומים פרקטיים אפשריים: שיפור ביצועים באימון מסווגים לביעות מולטימודליות עם חוסר איזון בין מודלים שונים.

פרטי מאמר:

[לינק למאמר: זמן להורדה](#).

[לינק לקוד: זמן-caan](#).

פורסם בתאריך: 20.10.2021, בארכ'יב.

הוגג בכנסו: NeurIPS 2020.

תחומי מאמר:

- מסויימים לבעיות מולטימודליות.
- שיטות רגולריזציה.

כליים מתמטיים, מושגים וסימונים:

- אנטרופיה פונקציונלית (FE).
- אינפורמציה פישר פונקציונלית.
- אי שוויונות לוגו של סובולב ושל פאונקרה.
- טנדורייזציה במרחב הסתברות מכפליים (product probability spaces).

תמצית מאמר:

המאמר מציע שיטה להתמודד עם הסטייה בכיוון של תת-קבוצה של מודים בתהיליך אימון על בעיות סיוג מולטימודליות. כאשר זה קורה המסוווג עלול להיות מוטה כלפי תת-קבוצה של המודים ולהתעלם (להתחשב פחות) מהמודים האחרים. למשל ניקח לדוגמה DATA סט הנקרא Colored MNIST שסט האימון וסט הולידייצה שלו מכילים תമונות צבעוניות של הספרות, והטסס סט מכיל תמונות בגווני אפור. אם מאמנים רשת ניירונים עם ליאו רגיל במטרה לסwoג את הספרה, ביצועה (דיקוק) על הטסס סט סופגים ירידיה שימושותית יחסית לביצועים על סט האימון ועל סט הולידייצה. הסיבה לכך היא שהמסוווג למד להתחשב בעיקר במצבו של תמונה (המוד הראשון) ומתעלם לרוב מהצורה של הספרה (המוד השני). עוד דוגמא לכך, היא משימת "מענה על שאלות ויזואליות" (מולטימודליות בזרה אינהרנטיות), כאשר המסוווג עלול לבחור להתמקד רק במציאות תשובה "הגיגונית" לשאליה ויתעלם מהאינפורמציה היזואלית. בשביל להתמודד עם סוגיה זו המחברים מציעים להוסיף לפונקציית לוס איבר שמנסה "להカリיח כל מוד לתנורום" לסיוג הסופי. איבר זה מבוסס על אנטרופיה פונקציונלית (FE), שבעזרתה ניתן לאמוד את התרומה של כל מוד לתוצאה של המסוווג.

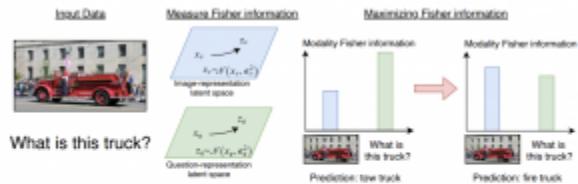


Figure 1: We illustrate our approach. In the visual question answering task, we are given a question about an image. Thus, we can partition our input into two modalities: a textual modality, and a visual modality. We measure the modalities' functional Fisher information by evaluating the sensitivity of the prediction by perturbing each modality. We maximize the functional Fisher information by incorporating it into our loss as a regularization term. Our results show that our regularization permits higher utilization of the visual modality.

במילים אחרות FE משערת את "מידת השתנות המומוצעת של התפלגות הפלט של המסוג עבור פרט/orbzיות (הרעשה) של הקלט" (במרחבי הייצוג ולא במרחבים המקוריים!!) תחת התפלגות מסוימת על הפרט/orbzיות (המאמר משתמש בתפלגות גאוסית באופן טבעי). ככל שמידת השתנות זו קטנה יותר עבור סיווג נתון (כלומר התפלגות של פלט המסוג איננה משתנה בהרבה עבור נקודות בסביבה יחסית רוחקה של הקלט – במרחב הייצוג), המסוג נוטה "לא לנצל מידע על לפחות חלק מהמודים באופן מיטבי".

בנוסף מכיוון FE קשורה לשונות של מידת השתנות של התפלגות הפלט, המאמר מציע להוסיף את איבר המשערך שונות זו במקום האיבר של FE. למעשה, ככל שהשינוי של מידת השתנות קטן יותר, התפלגות הפלט פחות תליה בקלט שזה כמובן לא רצוי. מעניין שאיבר רגולרייזציה בזורה של שונות מושג שייפור מסוים בביטויים עבור כמה מושימות.

רעיון בסיסי:

הרעיון הבסיסי של המאמר הוא להוסיף איבר המשערך FE לפונקציית לוס רגילה (קרוס-אנטרופי). מידת ההשתנות של פלט המסוג עבור פרט/orbzיות (הרעשה) של דוגמא נתונה מוגדרת במאמר בטור קרוס אנטרופי בין לבין התפלגות הפלט שלה (הקלט הנקי והקלט מורעש).

תקציר מאמר:

הבעיה העיקרית עם אנטרופיה הפונקציונלית נעוצה בעובדה שלא ניתן לחשב אותה באופן ישיר, אלא רק באמצעות דגימות של האנטגרנד. בנוסף לכך FE מכילה איבר לוג של אינטגרל התוחלת ששיעורכו ע"י" דגימות עלול להיות מאוד לא מדויק.

איןפורמצית פישר פונקציונלית: המאמר מציע להשתמש בהתאם מלעיל ל- FE (הנובע מאי שוויונות לוג של סובולב) ע"י" איןפורמצית פישר פונקציונלית (FFI), המהווה הכללה של איןפורמצית פישר קלסית. כמו שאתם אולי זכרים, איןפורמצית פישר רגילה (FI) מוגדרת עד כמה מידע יש בדגימות של משתנה מקרי X המפוגג עם פונקציית התפלגות f התלויה בפרמטר θ , על הפרמטר זהה. FI מוגדרת בתור תוחלת f' של הנגזרת הריבועית לפי θ של הלוג של f. עבור ערך פרמטר נתון FI משערת את "מידת הקשר" בין ערכו של הפרמטר לבין הדגימות של המשתנה המקרים X. ככל ש- FI יותר גבוהה ניתן לשערת את בפרמטר ביותר דיק (אפשר להסיק זאת גם מאי השווון של ראו-קרמר – הדיק נמדד שם ע"י" שגיאת שערוך ריבועית המומוצעת).

از הכללה של FFI מتباطאת בכך שמקיפים את האנטגרנד בפונקציית התפלגות נוספת (במקרה שלנו זה פונקציית התפלגות על הפרט/orbzיות של נקודות הדאטה (z_{per_f})). זה מאפשר להחליף את פונקציית הצפיפות f המופיעה בביטוי המקורי של FI בכל פונקציה אי שלילית (במקרה שלנו הפונקציה זו היא קרוס-אנטרופי בין

התפלגיות של הפלטימ (z)ce_f). בדומה ל FFI, גם מושער את מידת ההשתנות של (z)ce_f כאשר הפרמטר z מתפלג עם f_per(z).

חסם לוג של סובולב על FFI: חסם סובולב מאפשר לחסום את FE מלמעלה ע"י האינטגרל המכיל ממנו שמהונת בו הינו נורמה של הגראדיאנט הריבועי של קרוס-אנטרכופי (z) C בין התפלגיות של הפלטימ (z)ce_f כאשר המונע הוא (z) C עצמו. שניהם ניתנים לחישוב בצורה מפורשת וניתן להכניסם כמו שהם לפונקציית לוא.

טנזריזציה של FE לביעות מולטימודליות: המאמר מצין שנקודה במרחב הלטנטי של ממשימה מולטימודליות עשויה להיות מורכבת מכמה מודליות מהרחבים הלטנטים השונים. למשל במקהה של משימת מענה על שאלות ויזואליות, המרחב הלטנטי הוא בעצם מרחב פרויקט של שלושה מרחבים: "צוג התמונה", "צוג השאלה" ו"צוג התשובה". המאמר מראה שבמקרים כאלה ניתן לתאר את FE של נקודת DATAה בתור הסכום של האיברים שכל אחד מהם זה FE המוצע של כל מוד z, כאשר הממוצע המוחשב על מרחב הפרויקט של שאר המודלים (פרט ל i) עם פונקציות התפלגות המורכבות ממכפלה של (z)ce_f (פרט ל i). צוג זה נקרא טנזורייזציה והוא מאפשר לחשב חסם על FI של נקודת DATAה בצורה יחסית קלה. כרגע יש לנו את כל הכללים בשבייל לתאר את המבנה של פונקציית לוא המוצעת. לפני שנעבור לתייאור של פונקציית הלואណון קוצרות בצורה השניה של איבר רגולרייזציה המוצע במאמר קרי שונות של f_per(z) תחת (z)ce_f.

איבר רגולרייזציה בצורה של שונות של f_ce(z): ראשית, המאמר מצין שקיים קשר בין FFI לשונות של (z)ce_f כאשר תקף עבור ערכים קבועים של (z)ce_f. באופן אינטואיטיבי ככל ש FFI של פונקציית (z)g יותר גבוהה אז השונות נטה להיות גבוהה יותר כי שניים מתארים את מידת השתנות של (z)g תחת אותה מידת הסתרבות. בדומה ל FFI גם השונות לא ניתנת לחישוב בצורה מפורשת (רק ע"י דגימות) ואז המאמר משתמש באנו שוויון פואנקרה בשבייל להקל על החישוב. לבסוף מבצעים טנזורייזציה של הביטוי המתkeletal (משתמשים במשפט אפרון-שטרן) כדי לקבל את הביטוי הסופי.

מבנה של פונקציית לוא: מכיוון שהמטרה שלנו היא למקסם את FI, איבר רגולרייזציה המתווסף לפונקציה לוא מכיל את ההפכית של החסם על FFI (או שונות) ומחרבים אותו לוא קרוס-אנטרכופי סטנדרטי. זה כל הסיפור אפשר להתחיל את האימון.

הישגי מאמר:

המאמר מדווח על השיפור בדיק על 4 נתונים סטיטים מולטימודליים מול כמה גישות עדכניות.

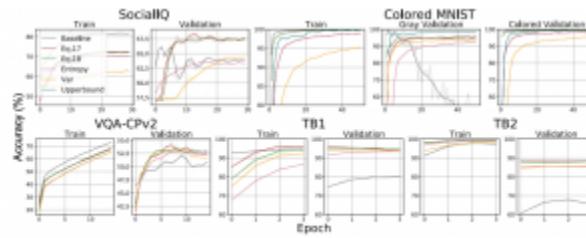


Figure 3: Training process with and without regularization. We note that generalization significantly improves when using our proposed regularization.

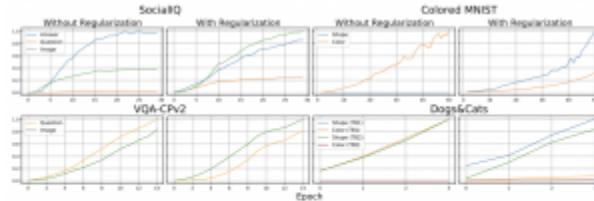


Figure 2: Proportions of the Fisher information values during training for SocialIQ, Colored MNIST, VQA-CPv2 and Dogs&Cats. Using our proposed regularization brings the modalities Fisher information value closer than training without our regularization, a desired property in multi-modal learning. In ColoredMNIST, we observe that training a model with our regularization, the prediction is based on both the shape and the color. Unlike, a model trained without our regularization which makes predictions based on the color only.

דעתהוטים:

- Dogs and Cats
- (הבנת מצבים בOIDAO) SocialIQ
- (מענה על שאלות ויזואליות) Colored MNIST
- VQA-CPv2

ג.ב.

הרעיון של המאמר די מוגבל, אך לי זמן להבין אותו כי המתמטיקה במאמר די קשוחה. מצפה לראות שימושים של טכניקה זו למגוון רחב של בעיות מולטימודליות (למשל בלמידת באמצעות חיזוקים عمוקה).

Review 39: Supermasks in Superposition

פינת הסוקר:

המלצת קריאה ממיק: מומלץ מאוד - יש במאמר שני רעיונות מגניבים.

בהירות כתיבה: בינונית פלאס.

רמת היכרות עם כלים מתמטיים וטכניקות של DL/ML הנדרשים להבנת מאמר: הבנה בסיסית בתחום למידה מתמשכת (continual learning), בלמידה מתמשכת וברשותות הופפייד.

ישומים פרקטיים אפשריים: בניית רשת נירונית גדולה עם משקלים קבועים המשמשת לביצוע משימות רבות (דומות באופי).

פרטי מאמר:

lienק למאמר: [זמן להורדה](#).

lienק לקוד: [זמן-can](#).

פורסם בתאריך: 22.10.20, בארכ'יב.

הוגג בכנס: NeurIPS 2020.

תחומי מאמר:

- שיטות למידה מתמשכת (continual learning) עם רשותות נוירונים.
- למידת משימות מרובות (multi-task learning) עם רשותות נוירונים.

כליים מתמטיים, מושגים וסימונים:

- מסכות בינהיות על משלבים ברשותות נוירונים.
- שכחה קטסטרופלית ברשותות נוירונים
- רשותות הופFIELD (HN).
- אנטרופיה (זה המושג המרכזי שעליו המאמר בניו).

תמצית מאמר:

המאמר מציע שיטה אימון SupSup של רשת נוירונים גדולה (נקרא לה רשת בסיס), המKENNA לה יכולה לבצע כמה משימות שונות. המשקלים של רשת הבסיס הינם קבועים לכל המשימות. בעצם לוקחים רשת, מעתחלים את משקליה באופן רנדומלי ומשתמשים בהואת רשת לחיזוי עבור משימות שונות. הדרך לבצע זאת זה למוד סט של מסכות בינהיות נפרד (0 או 1) לכל משימה ועבור כל משימה "להלביש את סט המסכות" שלא על רשת הבסיס בזמן אינפרנס. בעצם מסכה בינהית כזו מדליקה או מכבה קשרים בין נוירונים שונים ברשת. בדרך זו מתגברים על תופעת השכחת הקטסטרופלית שלולה להתறחש אם מאנים (מכילים) רשת עבור משימה חדשה.

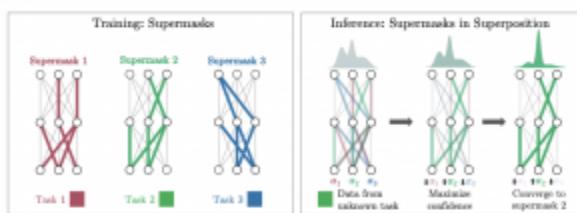


Figure 1: (left) During training SupSup learns a separate supermask (subnetwork) for each task. (right) At inference time, SupSup can infer task identity by superimposing all supermasks, each weighted by an α_t , and using gradients to maximize confidence.

המאמר מגדר 4 תרחישים טיפוסיים של למידה מתמשכת ומציע שיטה לאמן רשת בסיס אחת לביצוע של משימות מרבות עברו כל אחד מהם:

- המשימות ידועות (מצוות) גם במהלך האימון וגם במהלך האינפרנס (כלומר כל פעם שאנו מקבלים משימה אני יודעים איזו משימה זו): תרחיש זה מסומן כ-GG.
- המשימות ידועות במהלך האימון ולא ידועות במהלך האינפרנס (עם לייבור משותפים). קרי אנחנו צריכים לנחש איזו משימה קיבלנו באינפרנס בשביל להבין לאיזו משימה באימון היא שייכת (איזו מסכה לבחור): מסומן ע"י GN.

- המשימות ידועות במהלך האימון ולא ידועות במהלך האינפරנס עם לייבלים שונים. כלומר שכבת היציאה של הרשת צריכה להיות בגודל סכום של מספר הליבלים עבור כל המשימות: N_{NN} .
- המשימות לא ידועות לא בזמן האימון ולא בזמן האינפරנס (הליבלים חיברים להוות משתפים כאן). כאן באימון אנו מקבלים משימה ולא ידועים איזו משימה קיבלנו. לעומת כל פעם אנו צריכים להחליט האם להשתמש במסכה הקיימת או לאמן מסכה חדשה. דבר דומה קורה גם באינפראנס: NN .

ריעונות בסיסיים:

יש כמה ריעונות מעוניינים במאמר. ניתן להפריד אותם בגודל לשני סוגים:

סוג 1: שיטות לפתרון לכל התרחישים, המתוארים מעלה, של בעיות הלמידה המתמשכת:

- **משימות GG:** לבחור מסכה המתאימה עבור משימת אינפראנס.
- **משימות GN ו-GN_u:** המאמר מציע לתאר את המסכה עבור משימת אינפראנס כצירוף לינארי של כל המסכות שאומנו ולחפש את המקדים ע"י מציאת המקדם שהעלייה בו מביאה לירידה הכי גדולה בתנטרופיה של פלט הרשות עבור המשימה. כאמור המקדם, שהגדריאנט של האנטרופיה השילית (מכפלת ב- -1) של רשות הבסיס לפיו, הינה מקסימלית. זה מאפשר לא להריץ את הרשות עבור כל מסכות בזמן אינפראנס (אם יש מאות רבות או אלפי מסכות ודאות טיטים גדולים, זה יכול להיות ממשמעתי). במקום זאת מרכיבים את המשימה ברשות פעם אחת עם ממוצע של כל המשקלים (הם קוראים לזה One-Shot) ומחשבים את הגדריאנט. נציין ש- One-Shot מבוסס על גרדיאנט אחד בלבד של אנטרופיית רשות הבסיס שהוא פונקציה לא קווארה ביחס למקדמים של המסכות. עובדה זו עלולה להביא לבחירה של מסכה לא נכונה. בשביל להתגבר על בעיה זו המחברים הציעו לבחור את המקדם בתהילר איטרטיבי. כל פעם מאפסים חצי מהמקדים עם גרדיאנטים הכי נמוכים עד שנשאים עם מסכה אחת.
- **משימות NN:** עושים משהו דומה למתואר בסעיף הקודם אך אם לא נמצא מקדם שהשינו בו מביא לעלייה משמעותית בתנטרופיה (מחשבים softmax על כל הגדריאנטים), מאמנים מסכה חדשה. אחרת לוקחים את המקדם עבורו התקבל המקסימום ומשתמשים במסכה שלו. דרך אגב הם לא ציינו אופציה פשוטה נוספת לפתרון: במקום לאמן מסכה חדשה (במקרה שציריך) אפשר לאמן מקדים שיכולים לקטגוריות "לא קיימות". המאמר מנצל את ערכי המירונים האלה בשביל להחליפן את ממד הנגזרת לפי המקדים של צירוף לינארי של המסכה.

נוסף המחברים מציעים איזה טרייך נחמד התורם משמעותית לשיפור ביצועים. הם מօיפים "לייבלים מלאכותיים" למשימה כזכור ניירונים נוספים לשכבה האחורה של רשות הבסיס. למשל אם מאמנים מסוג MNIST עם 10 קלאסים אז השכבה האחורה תכלול, נגד, 100 ניירונים כאשר 90 הנירונים השווים ל- 10 الرجال שיכולים לקטגוריות "לא קיימות". המאמר מנצל את ערכי המירונים האלה בש سبيل להחליפן את ממד הנגזרת לפי המקדים של צירוף לינארי של המסכה.

Table 1: Overview of different Continual Learning scenarios. We suggest scenario names that provide an intuitive understanding of the variations in training, inference, and evaluation, while allowing a full coverage of the scenarios previously defined in [49] and [55]. See text for more complete description.

Scenario	Description	Task space discrete or continuous?	Example methods / task names used
GG	Task Given during train and Given during inference	Either	PiNN [42], Hatchill [51], PSP [46], "Task learning" [55], "Task-IL" [49]
GIn	Task Given during train, Not inference: shared labels	Either	IPWC [23], S1 [16], "Domain learning" [33], "Domain-IL" [10]
GIno	Task Given during train, Not inference: unshared labels	Discrete only	"Class learning" [33], "Class-IL" [49]
MIn	Task Not given during train. Nor inference: shared labels	Either	BCD, "Continuous/discrete task update learning" [25]

סוג 2: שיטות לשמירה יעילה ולהפקת מסכה מתאימה למשימה (עבור משימות NN)

המאמר מציע לשומר מסכות בראשת הופFIELD (NH) כאשר שומרים את כל המסכות בתוכה ע"י ביצוע עדכון המשקלים שלה (של NH). בזמן האינפראנס מנסים לאייר את המסכה האופטימלית ע"י חיפוש מינימום של הסכם המשוקל של פונקציית האנרגיה של NH (הלוס הרגיל שלה) והאנטרופיה של המסכה בראשת הבסיס (עבור המשימה שבנידון).

פינת האינטואיציה: עכשו ניקח כל רעיון המוצע במאמר וננסה להבין את הרצינול מאחריו.

גרדיאנט של אנטרופיית שכבת פלט של רשט בסיס ביחס למקדמים של הקומבינציה הליניארית של המסכות:

קודם כל שהרשט יותר בטוחה בחיזויו שלה עבור דוגמא נתונה, האנטרופיה של שכבת הפלט שלה תלך ותרד. למשל לווקטור פלט [0.05, 0.9, 0.45] (הרשט "מממש בטוחה" בחיזויו שלה) יש אנטרופיה נמוכה הרבה יותר מוקטור [0.3, 0.3, 0.3] (הרשט לא "בטוחה"). אך אם גרדיאנט של אנטרופיות שלילית של הרשת הוא גבוה עלייה במקדם זה תוביל לעלייה באנטרופיה השילילת כלומר לרידה באנטרופיה. ההנחה כאן שאם הרשת פולטה חיזויים "בטוחים" עבור מסכה מסוימת אז כנראה שקיבלו משימה דומה זו שהמסכה הזאת אומנה. שימוש לבשאהןטרופיה מחושבת על כל סט האימון של המשימה שהופך את ההנחה הזאת לסבירה.

הוספה של לייבלים מלאכוטיים לשכבת הפלט:

זה רעיון שמאוד אהבתי - כאשר אנו מאמנים רשת עם יותר נוירונים בשכבת הפלט עבור משימה מסוימת הרשת לומדת לשים שם ערכאים שליליים גבוהים בערך המוחלט (ההופכים לאפס עם softmax לאחר מכן). אך אם באינפראנס אנחנו מקבלים שם ערכאים שהם לא שליליים גבוהים זה סימן שימושה זו אינה תואמת למשימה עלייה מסכה זו אומנה. אך במקומות להשתמש באנטרופיה המאמר מציע לחשב את הלוגריתם של סכום האקספוננטים של הערכים בניירונים המלאכוטיים של השכבה האחורה. אם יוצא ערך גבוה אז המסכה לא מתאימה למשימה. עם כל היפוי בReLU הזה יש לי תחושה שניית להציג תוצאה דומה דרך הכנסת טמפרטורה בסיגמאיד למשחק.

שמירה של מסכות בראשת הופFIELD:

בשביל לחסוך מקום אחסון של המסכות המאמר מציע לשומר אותם בראשת הופFIELD NH. NH זה בעצם מטריצה שמרתתיה לאחסן וקטוריים המורכבים מ- $\{1, -1\}$ בזורה חסינה נגד רוש. המסכות של הרשותות מורכבות מ- $\{1, -1\}$ אך עושים טרנספורמציה פשוטה בשביל להפוך אותם לפורמט של NH. כל פעם שרוצים לאחסן וקטור נוסף בNH מעדכנים את המטריצה שלה עם הווקטור הזה (יש כמה דרכים לעשות זאת - הם השתמשו בכלל עדכן של סטודוקי). אך איך קוראים מטריצת הזכרון הזאת? נגיד קיבלו גרסה מורעשת של ווקטור, מזינים אותו לפונקציות אנרגיה המוגדרות ע"י מטריצה זו (בגודל זה צורה ריבועית שלה) ומנסים להביא אותה למינימום. ניתן להוכיח המינימום מתקיים בנקודת השמורה כי קרובה לקלט המורעש.

אבל במקרה שלנו (זה עובד רק בתרחיש GN) אנחנו לא רק צריכים לאייר את הווקטור השמור הקרוב ביותר לקלט (שהוא תמיד ממוצע של כל המסכות) אלא למצוא מסכה שמביאה למינימום את האנטרופיה של פלט הרשת. אך הם הושיבו ללוס הרגיל של NH איבר המכיל אנטרופיה של הרשת. בעצם הלוס הינו צירוף לינארי של לוס NH רגיל עם מקדם העולה עם האיטרציות והlös של אנטרופיה היורד עם מספר האיטרציות. האינטואיציה כאן שבהתחלתה צזים בכיוון של המסכה הנכונה וכאשר אנחנו באיזור ממצאים תחילה רגיל של מינימיציה אנרגיה של NH.

ולסימן אני רוצה לציין כי המאמר משתמש בראשת בסיס די גודלה ומאוד overparameterized. זה מאפשר למצוא מסכות, המאפשרות הרבה מהמשקלים שלה, שניתן לאמן לביצוע של משימות שונות. דרך אגב הם לא ציינו איך מתבצע האימון של כל מסכה פרט לשימה (יש מגוון שיטות).

הישגי מאמר:

המאמר מוכיח את העליונות של השיטה שלהם בכל התרחישים המתוארים לעלה וגם מראים שהביצועים שלהם לא רחוקים מהביצועים האופטימליים של רשות הבסיס למשימה (כאשר הרשות מאוננת לכל משימה בנפרד מחדש). הם גם מראים חיסכון משמעותי במקומ איחסון לשיטות עם ביצועים דומים. בנוסף הם גם מראים שהגישה שלהם מאפשרת לאמן אלפי משימות על רשות בסיס אחת עם ביצועים הלא נופלים בהרבה מהביצועים המקסימליים.

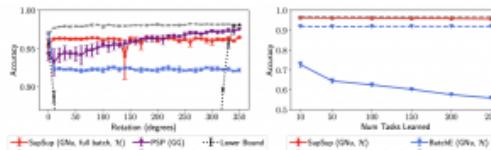


Figure 5: (left) Testing the FC 1024-1024 model on RotatedMNIST. SupSup uses **Binary** to infer task identity with a full batch as tasks are similar (differing by only 10 degrees). (right) The **One-Shot** algorithm can be used to infer task identity for BatchE [51]. Experiment conducted with FC 1024-1024 on PermutatedMNIST using an output size of 500, shown as mean and stddev over 3 runs.

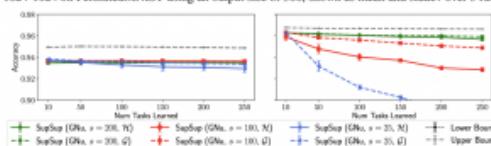


Figure 6: The effect of output size n on SupSup performance using the **One-Shot** algorithm. Results shown for PermutatedMNIST with LeNet 300-100 (left) and FC 1024-1024 (right).

Algorithm	Avg Top 1 Accuracy (%)	Bytes
Upper Bound	92.55	10222.81M
SupSup (GG)	89.58 88.68 86.37	195.18M 100.98M 65.50M
BatchE (GG)	81.50	124.99M
Single Model	-	102.23M

Figure 2: (left) **SplitImageNet** performance in Scenario GG. SupSup approaches upper bound performance with significantly fewer bytes. (right) **SplitCIFAR100** performance in Scenario GG shown as mean and standard deviation over 5 seed and splits. SupSup outperforms similar size baselines and benefits from *transfer*.

דאטאסתיטים:

SplitCIFAR100, SplitImageNet :GG.

.PermutatedMNIST, RotatedMNIST, SplitMNIST :GN

.PermutatedMNIST :NN

ג.ב.

המאמר מציע שיטה מבриיקה לאימון רשות אחת לביצוע של מספר גדול של משימות. עם זאת צריך לזכור כמה דברים:

1. המשימות שהם אימנו הם באוותה דרגת קושי (אני לא בטוח שהעוסק יעבד חלך אם המשימות היו בדרגות קושי שונות - אולי אז צריך לשחק עם מספר האחדים לכל מסכה בנפרד או משווה זהה).
2. המשימות שאומנו הן דומות מבחינה סמנטית. הם לאניסו לשלב דатаה סטיהם מודמיינים שונים.

3. המשימות שהם אימנו עליהם הן לא קשות ונשאלת השאלה אولي מבחן מקום אחסון עדיף לאמן רשות קטנה לכל משימה?

Review 40: Unsupervised Learning of Visual Features by Contrasting Cluster Assignments

פינט הוסף:

המלצת קריאה ממילק: חובה בטח לאוהבי למידת הייצוג.

בהירות כתיבה: בינוי פלאו.

רמת היכרות עם כלים מתמטיים וטכניקות של DL/ML הנדרשים להבנת מאמר: הבנה טובת בעקרונות הלווי המנוגד וידע טוב באופטימיזציה.

ישומים פרקטיים אפשריים: למידה ייצוגים חזקים על דאטסהטים לא מותאים עם תקציב חישוב מצומצם.

פרטי מאמר:

לינק למאמר: [זמן להודה](#).

לינק לקוד: [זמן-can](#).

פורסם בתאריך: 21.01.08, בארכיב.

הוזג בכנס: NeurIPS 2020.

תחומי מאמר:

- **למידת ייצוג ללא דאטסהט מותיג** (SSRL - self-supervised representation learning)
- **CLustering for deep representation learning** (CLIP) – SSRL מבוססת על טכניקות קליסטור

כלים מתמטיים, מושגים וסימונים:

- **מולטי-קרופ** – טכניקת אוגמנטציה המבוססת על קיחת פאטים'ים קטנים של תמונה ברזולוציות נמוכות שונות.

- האלגוריתם של סינקהורן-קנופ (Sinkhorn-Knopp) לפתרון בעיית הטרנספורט האופטימלי למידות הסתברות דיסקרטיות.

תמצית מאמר:

המאמר מציע שיטה למידת ייצוג על דאטסהט לא מותיג. רוב גישות המודרניות בתחום זהה (SSRL) מורכבות משני מרכיבים עיקריים:

- הלוֹס המנוגד (CL - contrastive loss): מסתמך על ההנחה שייצוגים של דוגמאות קרובות צריכים להיות קרובים, בזמן שייצוגים של דוגמאות לא קשורות (נבחנות רנדומלית בד"כ) צריכים להיות רחוקים.
- שיטה ליצירה של דוגמאות "דומות", קרי אוגמנטציה: בדרך כלל זוג דוגמאות קרובות (אקרה לזוגות האלה בהמשך זוגות חיובים או זוגות קרובים) נוצר ע"י הפעלה של שתי אוגמנטציות שונות על אותה דוגמא.

ນץין כי גישות SSRL המודרניות מסתמכות של השוואה של מספר גבוח מאוד של זוגות ייצוגים של דוגמאות שמנצירן כמות גדולה של זכרון ומשאבי ייבוד שימושותיים. דרישות אלו מקשות על שימוש של שיטות אלו בצורת אונליין (לטענת המאמר רוב שיטות SSRL היום מיושמת בצורת אונליין כדי הפטייע אותן). אך בוואו נדבר על החידושים שהמאמר הזה מציע:

- שיטת אימון VSwA: המאמר הנסקר מציע שיטה חדשה SSLR (הנקראת VSwA) העשויה להוריד גם את כמות החישובים וגם לצמצם את כמות הזיכרון הנדרשת. הרעיון העיקרי של המאמר הינו שינוי "ההגדירה של מושג הדמיון בין ייצוגי דוגמאות". למעשה המאמר "MAILZ" זוגות של הדוגמאות הקרובים "להשתיר" לאותם הקלאסטרים במרחב הייצוג במוקם להשווות את הייצוגים בצורה מפורשת (שיעור זה המוצג ע"י הקוד של דוגמא המחשב על סמך הבاطץ' שלו - אופן בנייתו פורט בהמשך). נץין ש-VSwA אינו דורש לשמר בזיכרון של דוגמאות שליליות שהופרדו אותו למיעמד טוב למימוש בצורת אונליין.
- שיטת אוגמנטציה מולטי-קרופ: המאמר מציע שיטת אוגמנטציה הנקראת מולטי-קרופ שמתחליה מהחישוב של שני "קרופים סטנדרטיים" cr_1 ו- cr_2 של תמונה x. לאחר מכן לוקחים "קרופים קטנים יותר" של cr_1 ו- cr_2 ב嚷ון רגולריות נמוכות ובונים מהם סט דוגמאות חיוביות עבור תמונה x. לטענת המאמר שיטה זו מקטינה את כמות החישובים הנדרשת תוך שמירה על הביצועים.

הסבר של רעיונות בסיסיים:

עכשו ננסה להבין מה פונקציית המטרה L שבילבה של שיטת VSwA. פונקציית L מוגדרת באופן הבא (לכל דוגמא בباء'ך):

- בונים מספר אוגמנטציות לדוגמא x עם מולטי-קרופ או כל גישה אחרת.
- מרכיבים מאוגמנטציות אלו זוגות של דוגמאות.
- בונים וקטורי ייצוג z לכל הדוגמאות שנבנו.
- לכל זוג וקטורי ייצוג (z_1, z_2) מחשבים את הקודים שלהם q_1 ו- q_2 .
- מחשבים את סכום הדמיונות s בין z_1 ו- z_2 לבין z_2 ול q_1 .
- מחשבים את הסכום x_L של כל הזוגות של הדוגמאות החיוביות של דוגמא x.

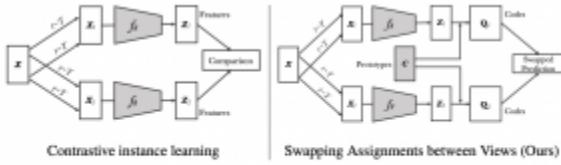


Figure 1: **Contrastive instance learning (left) vs. SwAV (right).** In contrastive learning methods applied to instance classification, the features from different transformations of the same images are compared directly to each other. In SwAV, we first obtain “codes” by assigning features to prototype vectors. We then solve a “swapped” prediction problem wherein the codes obtained from one data augmented view are predicted using the other view. Thus, SwAV does not directly compare image features. Prototype vectors are learned along with the ConvNet parameters by backpropagation.

פינט האינטואיציה:

למעשה תהליך אימון זה “مالץ” וקטורי ייצוג של דוגמא להכיל מידע על הקוד של הדוגמאות הקרובות. בוצרה לא פורמלית ניתן לומר שאנו מנסים למקסם את “המידע הדדי” בין הייצוגים של הדוגמאות שזה המטרה העיקרית של האימון עם הלווי המונגד CL. דרך אגב השם של השיטה נובע מהפעולה שחלוף (swap) שמבצעים בין הייצוגים ובין הקודים של דוגמאות קרובות באימון.

השאלה האחרון שטרם התיחסנו אליה הינה מבנה של פונקציית loss בין ייצוג z לקוד q ?

מבנה של פונקציה loss בין ייצוג z לקוד q (של דוגמאות הקרובות): אם אתם זוכרים הקוד q ניתן לפרש כוקטור הסתברויות שיר לקלסלטים. למעשה אנו רוצים שהקוד q ישיקף בצורה כמה שייתר טובה את המרחקים של z מהפרוטוטייפים \mathbf{c}_i שניתן לראות אותם בתור מרכזים (סנטרואידים) של קלסלטים של ייצוגים. אז קודם כל אחד בונים את וקטור המרחקים המנורמליים $-z$ לכל \mathbf{c}_i . מרחק זה מחושב אקספוננט של המכפלה הפנימית בין z ל \mathbf{c}_i . בסופו לוקחים את וקטור המרחקים ומנורמים אותו. לאחר מכן מחשבים את קרום אנטרופי בין q לווקטור מרחקים מנורמל שחייבנו. את הפונקציה זו אנו ממקסמים ביחס ל “יצוגים z וביחס לפוטוטייפים \mathbf{c} .

פינט האינטואיציה:

שיםו לב על הדמיון של המרחק בין וקטור הייצוג z ל- \mathbf{c}_i לביטוי של החוב המונגד CL. זה לא מקרי - אתם זוכרים שלhalb דיל מישיות מבודדות CL קלאסי, אין לנו כאן דוגמאות שליליות בצורה מפורשת. אז מה שמשחיק כאן את תפקיד “הדוגמאות השליליות” זה מרכזי הקלסלטים שרחוקים מ z . כמובן הם מאלצים ייצוגים של דוגמאות חייבות להיות רחוקים בצורה כמה שייתר דומה מכל הקלסלטים השליליים וקרובים באותה מידה מהקלסלטים החיוביים. לדעתך זה הנקודה הכי משמעותית במאמר (!!).

הסביר על בניית קוד q של ייצוג z : הקוד q של וקטור ייצוג z ממתאר את “רמת קרבתו” של z ל K וקטורי פרוטוטייפ \mathbf{c}_i . וקטור \mathbf{c}_i “מיצג” את הקלסלטר i . קוד של דוגמא (וגם של כל האוגמנטיות שלה) מחושב על סמך באטץ’ בודד בלבד (!!). אפשר להגיד שהקוד q מייצג את ההסתברויות שיר של וקטור הייצוג z של הדוגמא נתונה לקלסלטים המיאציגים עיי' וקטורי \mathbf{c}_i .

מטריצה Q המכילה את הקודים של כל הדוגמאות מהבאטץ’ הינה פתרון של בעיית אופטימיזציה לינארית עם איבר רגולרייזציה השווה לאנטרופיה הכלולית של Q (עם מקדם קטן). פונקציה מטרה זו מנסה למקסם את הדמיון הכלול בין וקטורי ייצוג של הדוגמאות בבאטץ’ לפרטוטייפים \mathbf{c}_i (כלומר לפחות את הקודים בצורה המשקפת את יחס המרחקים בין ייצוג הדוגמא למרכז הקלסלטים השונים). שימו לב שב下さいית אופטימיזציה זו מזכירה בוצרתת את בעיית הטרנספורט האופטימלי בין מידות הסתברות דיסקרטיות (האחדות) המוגדרות על שני דאטסהטים. את התפקיד של דאטסהטים כן משקפים הפרטוטייפים \mathbf{c}_i וקטורי הייצוג z של כל הדוגמאות בבאטץ’. המטרה כאן זה למצוא את האופטימלי שבו ניתן “להעביר את המשה ההסתברותית מוקטוריו \mathbf{c}_i וקטורי z (מצין שפונקציית המרחק שיש בהגדלה של הטרנספורט האופטימלי הינה פרופורצionalית במקורה שלו)

למרחוק בין z ל- c). למעשה אנו מנסים למצוא מטריצה Q האי שלילית, שайיר (k,j) שלה מגדיר את המסה ההסתברותית המועברות מוקטור k_z לוקטור z_c, ככלור הסתברות השיר של k_z לקלستر של z_c. מכיוון שאנו רוצים שאותו מספר דוגמאות "שוויר" לכל קלסטר, מוסיפים אילוץ על סכום השורות וסכום העמודות של Q. בעיה זו פותרים בעזרת אלגוריתם איטרטיבי של [סינקhorן-קונפ.](#).

הסבר על מושגים חשובים במאמר:

שיטת אימון של גישות SSRL המודרניות: בדרך כלל בזמן האימון של SSRL לכל זוג של דוגמאות קרובות בונים מספר גדול של זוגות רנדומליים (אקרא לזוגות רוחקים או זוגות שליליים). כאן פונקציית המטרה od_F (שממקסימים אותה) הינה יחס בין אקספוננט של דמיון של "הזוג הקרוב" (בין הייצוגים שלהם) לסכום הדמיונות בין כל הזוגות שליליים. למשל בשיטת [SimCLR](#) כל באטץ מרכיב-m-N זוגות של דוגמאות קרובות (אגומנטציה של הדוגמא) המהווים את הזוגות החיבים כאשר עברו דוגמא נתונה, כל הדוגמאות פרט ל'בת הזוג' שלה נחשבות לדוגמא שלילית עבורה. פונקציה המטרה לכל באטץ הינה סכום של פונקציות המטרה של כל N2 דוגמאות של הבאטץ'.

בנק של ייצוג דוגמאות שליליות: ידוע שהגדלת מספר הזוגות השליליים לכל זוג חיובי באימון תורמת לעוצמת הייצוג של הדאטה. כתוצאה לכך משתמשים באטצים מאד גדולים (עשרות אלפי דוגמאות) שדרוש משאבי זכרון גדולים, כח עיבוד רב (צריך לחשב את הייצוג של عشرות אלפי דוגמאות מהבאטץ'). כדי להקטין את כוח העיבוד הנדרש הוצע ([MoCo](#)) "בנק הדוגמאות שליליות" מהבאטצים הקודמים המכיל את הייצוגים של הדוגמאות מכמה הבאטצים הקודמים. כל פעם דוגמים ממש ייצוגים של דוגמאות שליליות ומוסיפים את זה לייצוגים שליליים מהבאטץ הנוכחי. צריך לזכור שגישה זו כרוכה בהקצת משאבי אחסון נוספים לשימרת בנק זה.

הישגי מאמר:

המאמר מראה Sh-VaSVAW משולב עם מולטי-קרופ מצלה ליצור ייצוגים יותר חזקים משליטה בנית ייצוג רבות עבור מספר משימות. ההשוואה בדרכ הסטנדרטיבית: הוספה של שכבה לינארית לרשות הבונה ייצוג (עם משקלים מוקפאים) ובcheinת ביצועיה על משימה מסוימת. קודם כל הם הראה שייצוג שנבנה באמצעות Sh-VaSVAW מציג ביצועים יותר טובים על נתונים [Places205](#), [iNaturalist2018](#), [VOC07](#) ו- [Planes205](#) מהייצוגים הנבנים על ImageNet מתויג (!!) גם על משימת סיווג ועל משימת זיהוי אובייקטים. בנוסף הם הראו שהיצוגים שלהם משיגים ביצועים יותר טובים מבחינת Top1/Top5 (локחים 1/5 דוגמאות היכי קרובות מבחינת הייצוג ומחשבים כמה מתוכם שייכים לאוთה קטgorיה) מ- [MoCov2](#) ו- [SimCLR](#). מזכיר שלהבדיל מ-2, אין צורך לשמור של בנק דוגמאות שליליות ב- Sh-VaSVAW. הם גם הראה את עלילנותה של Sh-VaSVAW במשימות semi-supervised על שיטות כמו AUDA ו- FixMatch. וזה רק חלק קטן מכל השוואות שהם עשו - הם באמת עשו עבודה מרשימה בהיבט זהה.

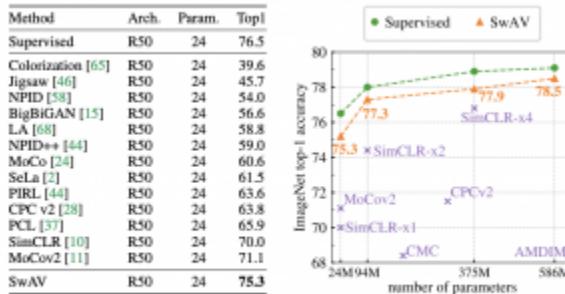


Figure 2: **Linear classification on ImageNet.** Top-1 accuracy for linear models trained on frozen features from different self-supervised methods. (left) Performance with a standard ResNet-50. (right) Performance as we multiply the width of a ResNet-50 by a factor $\times 2$, $\times 4$, and $\times 5$.

Table 1: Semi-supervised learning on ImageNet with a ResNet-50. We finetune the model with 1% and 10% labels and report top-1 and top-5 accuracies. *: uses RandAugment [12].

Method	1% labels		10% labels	
	Top-1	Top-5	Top-1	Top-5
Supervised	25.4	48.4	56.4	80.4
<i>Methods using label-propagation</i>	-	-	68.8*	88.5*
UDA [60]	-	-	71.5*	89.1*
FixMatch [51]	-	-	-	-
<i>Methods using self-supervision only</i>	-	-	-	-
PERL [44]	30.7	57.2	60.4	83.8
PCL [37]	-	75.6	-	86.2
SimCLR [10]	48.3	75.5	65.6	87.8
SwAV	53.9	78.5	70.2	89.9

ג.ב.

מאמר ממש מגניב עם רעיון מתקדם המשלב תובנות רבות מגוון שיטת SSRL. הם גם טרחו להשוות את הביצועים של השיטה שלהם מול מגוון רחב של אלגוריתמים, משימות, DATA סטים וקונפיגורציות שונות בהחלה מרשימים. בקיצור המלצת קריאה לוחתת ממנה:

שנקרא:

Review 41: Improving GAN Training with Probability Ratio Clipping and Sample Reweighting

פינית הוסף:

המלצת קריאה ממילא: מומלץ אך לא חובה לאלו שרצו להתעמק בשיטות אימון של GANs.

בהירות כתיבה: ביןונית פלאו.

רמת היכרות עם **כלים מתמטיים וטכניקות של ML/DL הנדרשים להבנת מאמר:** הבנה טובה ווורשטיין גאן וכל מה הקשור אליו, הכרה בסיסית בשיטות מעולם הסטטיסטייה כמו importance sampling, reinforcement learning (Reinforcement learning) .

ישומיים פרקטיים אפשריים: אימון גאן משופר במגוון תרחישים

פרטי מאמר:

[لينك لمقالة: זמן להורדה.](#)

[لينك לקוד: זמן CAN.](#)

פורסם בתאריך: 30.10.2020, בארכיב.

הוזג בכנס: NeurIPS 2020

תחומי מאמר:

- ganims.
- שיטות אימון של ganims.

כליים מתמטיים, מושגים וסימונים:

- וירשתין GAN (WGAN).
- מרחיק ווירשתין (WD).
- פונקציית לפשייז.
- שיטות וריאציניות לביעות אופטימיזציה בתחום הרשותות הגנרטיביות כמו N-GAN.
- גישהת מתרורת למידת החיזוק (RL): אופטימיזציה של פוליסי (PO - Policy Optimization) דרך פתרון של בעית אופטימיזציה עם פונקציית מטרה חלופית - surrogate.
- שיטות דגימה: Importance Sampling (IM).
- מרחיקים בין מידות הסתברות: מרחק KL ומרחיק KL ההפוך.
- אלגוריתמים של SOTA expectation-maximization (EM) Expectation-Maximization.

תמצית מאמר:

אתם בטח יודעים של מרירות מאמץ מחקר אינטנסיביים בשנים האחרונות, האימון של GAN-ים עלול להוות משימה לא טריומיאלית עקב קושי במציאת איזון בין הגנרטור G לדיסקרמיןטור D . המאמר הנסקר מצין שביעיות אלו בולטות במיוחד בתחום גנרטוט טקסט עקב האופי הדיסקרטי של משימה זו (нациין שכרגע שיטות SOTA למשימות גנרטוט של טקסט אינם מבוססות על GAN-ים). כדי להתגבר על סוגיות אלו, מאמר הנסקר מציע שיטה לשיפור תהליכי האימון של GAN שבסיסת על שני רעיונות עיקריים:

- מניעה עדכנים גדולים מדי של הגנרטור G שעולמים לפגוע ביציבות של תהליכי האימון ולהוביל לאובדן של איזון בין G לדיסקרמיןטור D . איזון זה הינו חיוני להתקנות של תהליכי האימון של GAN ולפתרון איקוטי עבור בעית אופטימיזציה מינימקס Sh-GAN מנסה לפתור. נזכיר שתהליכי האימון של GAN הינו משחק סכום אפס כאשר G מאמין לגורם D להזחות את הדעתה הסינטטי Sh-G מיצר כדטה אמיתית (מסט האימון) ובתורו D מאמין להבחן בין דוגמאות Sh-G מיצר לאמתיות.
- משמעות המגנרטות עי G בתהליכי האימון של D . כאמור D מאמין להבחן בין דוגמאות אמיתיות (מאומן לתת ציון גבוהה) מסט האימון לבין דוגמאות המגנרטות עי G (מאומן לתת ציון נמוך). בתהליכי עדכן של D הדוגמאות של G באיכות טובה שמצוות "לעבוד יותר טוב על D " (בעל ציון גבוהה) מקבלות משקל גבוה ואילו דוגמאות של G ה"פחות אמיתיות" מבחינת D (בעל ציון נמוך) מקבלות משקל נמוך יותר. זה הופך את האימון של D לעיל יותר כי (לטענת המאמר) הוא לא מתבזבז על עדכנים על דוגמאות קלות מדי (האינטואיציה כאן אומרת שאם D משקיע מאמץ רב יותר בהתאם על דוגמאות איכותיות יותר, הוא יהיה מספיק חזק בשבייל להפגין ביצועים טובים גם על דוגמאות קלות יותר ב"צורה אוטומטית").

הערה: גישה זה מזכירה לי שיטות משפחת gradient boosting machines (GBM) ממושכות דוגמאות בהתאם ל"רמת הקושי" שלהם מבחינת המודל (בגדול עד כמה השערור של המודל מדויק).

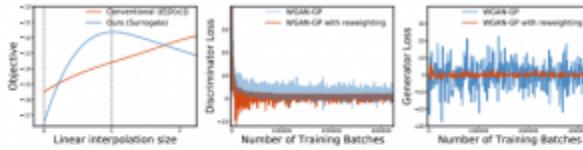


Figure 1: Illustration of the proposed approach for stabilizing GAN training. Results are from the CIFAR-10 experiment in Sec.4.1. **Left:** The conventional and surrogate objectives for generator training, as we interpolate between the initial generator parameters θ_{old} and the updated generator parameters θ_{new} , which we compute after one iteration of training. The θ_{new} obtains maximal surrogate objective. The surrogate objective curve starts decreasing after $x = 1$, showing the objective imposes a penalty for having too large of a generator update. In contrast, the conventional objective (for WGAN-GP) keeps increasing with larger generator updates. **Middle and right:** Discriminator and generator losses w/o and w/o sample re-weighting. WGAN-GP with our re-weighting plugged in shows lower variance in both discriminator and generator losses throughout training.

הסבר של רעיונות בסיסיים:

ורשטיין GAN: נקודת ההתחלה של המאמר זה WGAN, המודיפיקציה של ה-GAN המקורי, המשמשת במרקח ורשטיין (WD) כבסיס ל-D. כלומר G מאמן לגנרט דגימות בעלות מרחק ורשטיין נמוך מהדוגמאות מסט האימון. מרחק ורשטיין הינו מקרה פרטי של טרנספורט אופטימלי וכבר הסרתי על באחד הפוסטים שלו ([Learning to summarize from human feedback](#)).

היתרון הבולט של WGAN על GAN רגיל טמון ביכולת של D "להעביר גרדיאנטים" יותר יציבים ל-G גם במקרים כאשר D מצליח בקשות להבדיל בין הדגימות האמיתיות לדגימות המוגנרטות. זה קורה בגל של להבדיל מרחק Jensen-Shannon (JS) שהוא מנסה למצער ה-GAN הרגיל, WD הינו בעל אופי רציף יותר ולא מגע לרזיה (כמו מרחק JS) גם כאשר התפלגות הדגימות של G רוחקה מאוד מההתפלגות של הדאטה סט (המשוערכת ע"י D).

чисוב של מרחק ורשטיין לפי הגדרתו הינו משימה מאד קשה ובדרך כלל פוטרים את בעיית האופטימיזציה הדואלית שלה (שוויון רובינשטיין-קנטורוביץ'). הבעיה הדואלית הינה המקסום של הפרש התוחלות בין התפלגות של דатаה האמיתית לבין הדגימות המוגנרטות מעלה מרחב של פונקציות $\text{-}\lambda$ -ליפשיץ רציפה, מוכפלת ב $1/k$. פונקציה זו מודדת ע"י רשת נירונים כאשר נעשים טרייקים שונים, כמו קיצוץ משקלים או אילוצים על הנגזרת של הפונקציה כדי שהפונקציה המודולת תהיה $\text{-}\lambda$ -ליפשיץ רציפה). אז בעית אופטימיזציה ש-WGAN מנסה לפתור, הינה מקסום של הפרש התוחלות זה מעלה מרחב כל פונקציות $\text{-}\lambda$ -ליפשיץ רציפות f , מבחינת D. הגנרטור G מצדיו מנסה למצער אותו הפרש התוחלות המתואר לעיל (בעית מינימקס). אם נתבונן בפונקציית מטרה של WGAN ניתן לראות כי G מנסה למקסם את התוחלת של פונקציה $\text{-}\lambda$ -ליפשיץ f (על מרחב הדגימות שלו). ניתן למצוא דמיון בין בעית אופטימיזציה זו לבין אופטימיזציה של פוליסי בעולם של RL, כאשר פונקציה $\text{-}\lambda$ -ליפשיץ רציפה f משחקת תפקיד של גמול (reward) והתפלגות דגימות של G ניתנת לראות כפוליסי. דמיון זה, שזוהה בכמה מאמרם של השנים האחרונות, יונצל לבנייה של פונקציית מטרה חדשה ל-WGAN שהוצעה במאמר.

אחרי שהבנו מה זה WGAN ואת הקשר שלו לבעיות RL, בואו נתקדם בשינוי של פונקציית מטרה של WGAN המוצע ע"י המאמר. פתרונה יוביל למניעה של עדכנים גדולים של G ומשקל דגימות, המבוסס על ה"aicotas" שהן בעדכנים של D. לאור הקשר עם בעיות של אופטימיזציה של פוליסי כMO PPO ו-TRPO. שיטות אלה מחליפות את פונקציית המטרה הרגילה בפונקציה חלופית שמנסה לשפר את פונקציית הפוליסי $\text{-}_k F$. זה געשה ע"י מקסום התוחלת של פונקציית היתרון המוכפלת ביחס של $\text{-}_k F$ החדש ל- $\text{-}_k F$ הישנה תחת אילוץ שマーク KL בין $\text{-}_k F$ החדש לשנה חסום ע"י קבוע קטן (אילוץ זה מופיע לפחות פעם אחת רגולרייזציה בפונקציית המטרה). בדרך זו $\text{-}_k F$ החדש לומדת לתת הסתברויות גבוהות למצבים שבהם פונקציית היתרון מקבלת ערכים גבוהים ככלمر הגמול אחריו עדכן של $\text{-}_k P$ מקסימלי).

פונקציית המטרה של המאמר: המאמר מציע להחליף את פונקציית המטרה הסטנדרטית של WGAN בפונקציה imp_F המכילה הפרש של שני האיברים הבאים:

- איבר 1: התוחלת של פונקציה $\text{A-Lip}(\mathbf{f}, \mathbf{g})$ מעל מידת הסתברות עזר \mathbf{q} (שתייה בתפלגות הדגימות המוגנרטות \mathbf{g}_P וגם בפונקציית \mathbf{f} המודולת ע"י \mathbf{D} בצורה מפורשת ולא פרמטרית (!!!)).
- איבר 2: מרחק KL בין \mathbf{q} לבין \mathbf{g}_P .

המאמר מציע לאמן את WGAN ע"י מקסום של imp_F , כאשר הפרמטרים הם משקל החלטות של \mathbf{G} ו- \mathbf{D} . אם נזכיר בעובדה שמרחק KL הינו תמיד אי-שלילי, ניתן להבין שהמаксום של imp_F שקול למינימום של האיבר הראשון המינימיזציה של האיבר השני. אז ניתן לפרש את בעיית מינימום imp_F באופן הבא:

מקסום של תוחלת הציון הנitinן ע"י \mathbf{D} לתפלגות \mathbf{q} (האיבר הראשון) כאשר אנו מנסים לשמר את התפלגות הדגימות של \mathbf{G} קרובה ל- \mathbf{q} .

אימון של \mathbf{G} : מינימום של imp_F מבחינת הפרמטרים של \mathbf{G} , הינו מקרה קלאסי של בעיית אינפראנס וריציאוניות שמצוירה את בעית אופטימיזציה שאנו פותרים למשל ב-VAE- Variational AutoEncoder. הדרך הטבעית לפתרור אותה הינה להשתמש באלגוריתם EM קלאסי. בשלב E של EM, אנו מוצאים את התפלגות \mathbf{g} שהיא בצורה של מכפלה של אקספוננט של \mathbf{g}_P ושל \mathbf{f} (מנורמלת). שימוש לב שמה שיש מכפלה זו מהו משקל של \mathbf{g}_P , אשר הדגימות עם ציון של \mathbf{D} יותר גבוהה מקובלות הסתברות גבוהה יותר, זהה מה שרצינו מההתחלת.

השלב M של האלגוריתם הינו אופטימיזציה של imp_F על הפרמטרים של \mathbf{G} כאשר התפלגות \mathbf{q} נתונה (חוسبה בשלב E). זה למעשה מינימיזציה של האיבר השני, מרחק KL. וכך יש לנו בעיה כי \mathbf{q} זה בעצם פונקציה של \mathbf{g}_P הניתנת לצורה לא מפורשת ובשביל לשערק את מרחק KL נוצרת לדגום מ- \mathbf{q} שזה מאד לא טריויאלי. למזלנו ניתן להשתמש ב-BL הפור ולהפוך את האיבר זה לסקום של מינוס התוחלת של \mathbf{f} מעל \mathbf{g}_P ומרחק KL בין \mathbf{g}_P לבין \mathbf{g}_Q שעבור האיטרציה הקודמת לביין \mathbf{g}_P שאנו מנסים לאפטם (נוסחה 4 במאמר). בעצם אנו מנסים למינום את התוחלת של \mathbf{f} מעל \mathbf{g}_P אך לא רוצים להתרחק מדי מההתפלגות \mathbf{g}_P מהאייטרציה הקודמת. אם אTEX זוכרים את ההסבר שלי על PPO ועל TRPO, מיד תזהו את הדמיון. אז בדומה לשיטות אלו, המאמר מציע להחליף את פונקציית המטרה כאן בפונקציית מטרה חלופית המכילה המינוס של פונקציה \mathbf{f} ביחס בין \mathbf{g}_P הישן לחישוב \mathbf{g}_Q (!!). בנוסף הם מאלצים את \mathbf{g}_Q להיות קטן באופן מאולץ (מקצתים). אבל כאן יש לנו עוד בעיה. איך נחשב את היחס הזה על דוגימה של \mathbf{G} אם \mathbf{g}_P נתון לצורה לא מפורשת. כאן הם עושים טריק נחמד. בנוסף ל- \mathbf{D} של WGAN, הם מאמנים דיסקרימינטור בגיןarian $\text{ho}\mathbf{D}_P$ בשביל להבדיל בין הדגימות של \mathbf{G} לדגימות האמיטיות. ניתן להוכיח (עשוי זאת במאמר המקורי של GAN למשל) שעבור $\text{ho}\mathbf{D}_P$ אופטימלי ניתן לחשב את ערך של \mathbf{g}_P עבור הדגימה של הערך של $\text{ho}\mathbf{D}_P$ הדגימה זו. בדרכז זו ניתן לשערק את \mathbf{g}_Q עברו דוגימה נתונה.

אימון של \mathbf{D} : כאן אנו צריכים לאפטם רק את האיבר הראשון (התוחלת של \mathbf{f} מעל התפלגות \mathbf{q} נתון כאשר מופיעים את הפרמטרים של \mathbf{f}). כאן משתמשים כਮובן ב-Gradient Descent אבל נשאלת השאלה איך נחשב את הגדריאנט עבור הפרמטרים של \mathbf{f} אם אנחנו לא ידועים לדגום מ- \mathbf{q} . בשביל להtagבר על הקושי זהה הם משתמשים בטכנית קלאסית בסטטיסטיקה הנקראית MI תוך ניצול של הצורה של \mathbf{q} (מכפלה של אקספוננט של \mathbf{g}_P ושל \mathbf{f}). בתרור התפלגות proposal שדוגמים ממנה במקומם \mathbf{q} , הם לקחו את \mathbf{g}_P שקל לדגום ממנו. נציג התוחלת של הגדריאנט מעל \mathbf{q} של \mathbf{f} יצאת שווה לתוחלת מעל \mathbf{g}_P של המכפלה של \mathbf{f} באקספוננט של \mathbf{f} . כך אנו מושגים את המשקל הגבוה לדגימות בעלות ציון גבוה מ- \mathbf{D} משפיקות יותר חזק על העדכון של \mathbf{D} כאשר השפעה של דגימות עם ציון נמוך על עדכון של \mathbf{D} קטנה (!!).

Algorithm 1 GAN Training with Probability Ratio Clipping and Sampling Re-weighting

```

1: Initialize the generator  $p_\theta$ , the discriminator  $f_\phi$ , and the auxiliary binary classifier  $C$ 
2: for  $t \leftarrow 1$  to  $T$  do
3:   for certain number of steps do
4:     Update the discriminator  $f_\phi$  with sample re-weighting through Eqs.(7)-(8), and maintain  $f_\phi$ 
       to have upper-bounded Lipschitz constant through, e.g., gradient penalty [15].
5:   end for
6:   for certain number of steps do
7:     Finetune the real/fake binary classifier  $C$  (for 1 step)
8:     Estimate probability ratio  $r_1(\theta)$  using  $C$  through Eq.(6)
9:     Update the generator  $p_\theta$  with probability ratio clipping through Eq.(5)
10:   end for
11: end for

```

הישגי מאמר:

דומין של תמונות: המאמר מראה שהשיטה שלהם משפרת את איכות התמונות מבחינות Inception Score ו-Frechet Distance מול כמה GAN-ים וביניהם אלו המבוססים על הלווי WGAN עם טכניקות יזכוב אימון שונות וגם על כמה GAN-ים עם פונקציות לוס אחרת (לא בסגנון וסרשטיון). הם גם מראים שהם אכן מצליחים לייצב את האימון ועבור WGAN קלואס (השנות של גראידנטים נמכה יותר וההתכונות יותר מהירה). הניסויים נעשו בעיקר על CIFAR10.

דומין טקסטואלי: הם הצליחו לשפר את איכות הטקסט המוגנרט - ההשוואה נעשתה ע"י BLEU. מעניין שהם גם הצליחו לשפר את איכות ביצוע המשימה של "העברת סגנון" (Style Transfer) כאשר המטרה כאן לשנות את סגנון המשפט (למשל סנטימנט) תוך כדי שימור התוכן.

Length	MLE	SeqGAN [56]	LeakGAN [16]	RelGAN [35]	WGAN-GP [15]	Ours	Real
20	9.038	8.736	7.038	6.680	6.89	5.67	5.750
40	10.411	10.310	7.191	6.765	6.78	6.14	4.071

Table 2: Oracle negative log-likelihood scores (\downarrow) on synthetic data.

Method	BLEU-2 (\uparrow)	BLEU-3 (\uparrow)	BLEU-4 (\uparrow)	BLEU-5 (\uparrow)	NLL _{gen} (\downarrow)	Human (\uparrow)
MLE	0.768	0.473	0.240	0.126	2.392	-
LeakGAN [16]	0.826	0.645	0.437	0.272	2.356	-
RelGAN 100 [35]	0.811	0.705	0.501	0.319	2.482	-
RelGAN 1000 [35]	0.837	0.654	0.435	0.265	2.285	3.42 \pm 1.23
WGAN-GP [15]	0.872	0.636	0.379	0.220	2.209	-
Ours	0.905	0.692	0.470	0.322	2.265	3.59 \pm 1.12

Table 3: Results on EMNLP2017 WMT News. BLEU measures text quality and NLL_{gen} evaluates sample diversity. Results of previous text GAN models are from [35], where RelGAN (100) and RelGAN (1000) use different hyper-parameter for gumbel-softmax. Our approach uses the same gumbel-softmax hyper-parameter as RelGAN (1000).

Method	IS (\uparrow)	FID (\downarrow)
Real data	11.24 \pm 12.7.8	-
WGAN-GP (2017)	7.86 \pm .08	-
CT-GAN (2018)	8.12 \pm .12	-
SN-GANs (2018)	8.22 \pm .05	21.7 \pm .21
WGAN-ALP (2020)	8.34 \pm .06	12.96 \pm .35
SRNGAN (2020)	8.53 \pm .04	19.83
Ours (re-weighting only)	8.45 \pm .14	13.21 \pm .60
Ours (full)	8.69\pm.13	10.70\pm.10

Table 1: CIFAR-10 results. Our method is run 3 times for average and standard deviation.



Figure 2: Generated samples by WGAN-GP (top-left), CT-GAN (bottom-left), and ours (right).

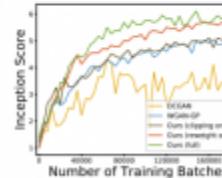
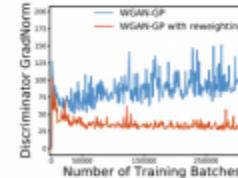


Figure 3: Left: Inception score on CIFAR-10 v.s. training batches (including both generator and discriminator batches). The DCGAN [39] architecture is used. Right: The gradient norms of discriminators on fake samples.



ג.ב.

אחד המאמרים היפים מבוחינת האלגוריתם המתב투א השילוב טכניקות מתחומיים שונים (לא צייני) בסקרה שהם מוכחים שהגישה שלהם מגדמת את התפלגות של G לכיוון של התפלגות הדאיה האמיתית. לבג'י הישימות של גישה זו חיברים לבחון אותה על דатаה סטיטים יותר מגוונים ועל משימות מורכבות יותר.

Review 42: Representation Learning via Invariant Causal Mechanisms

פינת הסוקרי:

המלצת קריאה ממ"ק: מומלץ לאוהבי למידת ייצוג, בעלי ידע בסיסי בתורת הסיבתיות.

בahirot כתיבה: בינוי פלוס.

רמת היכרות עם כלים מתמטיים וטכניקות של DL/ML הנדרשים להבנת מאמר: היכרות בסיסית עם כלים מלמדת ייצוג ומתרות הסיבתיות.

ישומים פרקטיים אפשריים: שיפור ביצועים לכל שיטת למידת ייצוג המבוססת NCE.

פרטי מאמר:

lienק למאמר: [זמן להורדה](#).

lienק לקוד: לא נמצא בארכ'יב.

פורסם בתאריך: 15.10.20, בארכ'יב.

הוזג בכנס: ICLR 2021 Poster

תחום מאמר:

- למידת ייצוג (representation learning).
- תורת הסיבתיות.

כלים מתמטיים, מושגים וİMונים:

- גרפ סיבתיות של מודל הסטברות.
- InfoNCE - Contrastive Predictive Coding
- לוֹס נִיגּוֹד - [NCE](#).
- מרחק KL בין התפלגות.
- עידון של משימה למידה (task refinement).

תמצית מאמר:

המאמר מציע שיטה (הנקראת RELIC) לבנייה של ייצוג של דатаה במרחב מממד נמוך. הרעיון מהו זה הכללה של InfoNCE ותभטה בהוספת איבר רגורייזציה לפונקציית הלוֹס שלה. איבר רגורייזציה זה נועד "לזודד" שההתפלגות הדמיונית בין ה"ייצוגים" איננו ריאנטית תחת אוגמנטציות שונות המופעלות על הדוגמאות האלו" (במאמר זה גם נקרא שינוי סגנון ואשתמש בשני המושגים האלה בהמשך הסקירה) ארכיב על כך בהמשך.

از בואו נבין מה התוספת הזאת תורמת לפונקציית לוֹס. קודם כל השיטות מבוססות, estimation - noise contrastive estimation, בניוות בצורה שגורמת לייצוגים של דוגמאות "קרובות" להיות קרובים גם כן (במרחב הייצוג). עברו

דומין התמונות קירבה מוגדרת כDMINON מבחינה סמנטיית/תוכן. פועלות אוגננטציה כמו הזזה, סיבוב או קרוף אין משפיעות על דמיון (קירבה) בין "יצוגים של תמונות באופן משמעתי". איבר רגולרייזציה המוצע במאמר "مالץ" את הייצוגים, בנוסף לתוכנה המתוארת מעלה, להיות אינו-ריאנטיים לשינויים לא סמנטיים "שאין להם השפעה על הקירבה" (קרי, שינוי סגנון). במילים אחרות בהינתן הייצוגים של תמונות בעלות קירבה מסוימת ביניהם (הקירבה יכולה להיות גבוהה או נמוכה), הייצוגים של תמונות אלו לאחר האוגננטציה "מאולצים לשומר על אותה הקירבה כמו התמונות המקוריות". זו תוספת משמעותית ללוס הריגל של שיטות מבוססות NCE כי היא "מאולצת" את הייצוגים "ליציג את התוכן של התמונה בלבד (!!)" עם כמה שפחות תלות בסגנון של התמונה. זה מוביל לייצוג יותר רלוונטי וקורלטיבי למשימות downstream (הקשריות לתוכן) -זו בעצם הנחת יסוד של המאמר.

רעיון בסיסי:

הראיון הבסיסי של המאמר בניו על 3 הנחות יסוד שמאפשרות להציג את תהליך של ייצור תמונה כgraf סיבתי.

תהליך ייצור תמונה:

1. התמונה נוצרת ממשטנה לטנסי של תוכן C ומשתנה לטנסי של סגנון S
2. המשטנים S ו- C הינם בלתי תלויים (התוכן לא תלוי בסגנון).
3. רק תוכן של תמונה רלוונטי למשימות downstream שעבורו הייצוג נבנה. סגנון של תמונה אינו רלוונטי למשימות אלו כלומר שינוי סגנון לא מושפעות על תוצאה מסוימת \hat{Y} . לדוגמה במשימת סיוג עם שני קלאסים (נגדי כלבים וחתולים), איברי גוף שונים של כלבים ושל חתולים מהווים תוכן כאשר רקע, תנאי תאורה, אופינים של עדשת מצלמה וכדומה מיוחדים לסגנון.

תחת הנחות אלו תוכן של תמונה מהו יציג טוב שלא עבור משימות downstream וכתוצאה לכך המטרה של למידת ייצוג תוקן של תמונה. במקרים אחרים, משתנה תוכן של תמונה X מכיל את כל המידע הרלוונטי לחיזוי, המבוצע במסגרת משימה \hat{Y} , והוא צריך להיות אינו-ריאנטי (לא משתנה) תחת כל שינויים כלשהם של סגנון.

הסבר קצר על מושגי יסוד במאמר:

אחד ממושגי היסוד במאמר זה שיטות ללמידה ייצוג מבוססות NCE - בואו מրענן בקצרה את הנושא הזה:

שיטות NCE: הנחת היסוד ב- NCE מתבססת על ההנחה שיצוג חזק של DATA בהכרח מסוגל להפריד בין זוגות של דוגמאות דומות לבין זוגות דוגמאות רנדומלית. בין השימושים של טכניקה זו אפשר להזכיר negative sampling שהשתמשו בו למשל ב-word2vec. ניתן להוכיח שעבור צורה מסוימת של NCE לוס (הנראת InfoNCE) כי ככל שלוס זה קטן יותר המידע הדדי בין הדוגמא במרחב המקורי לבין ייצוגה למרחב ממימד נמוך עולה (ציריך לציין שהמאמר הנסקר טוען שיש עבדות שטוענות שהביצועים של ייצוגים על משימות downstream יותר תלויה בארכיטקטורה של האנקודר ופחות קרובה למידע הדדי). זה כמובן מצביע על אובדן פחות אינפורמציה בין הדטה המקורי לבין ייצוגה כולם הייצוג יהיה פחות לosi ומיצג את הדטה בצורה יותר מלאה. חשוב לציין שהאימון מtbצע במרחב היצוג ולא במרחב המקורי כולם הלוס מחושב על היצוגים למרחב ממימד נמוך. לוס NCE זה בעצם לוקח זוג דוגמאות קרגבות והרבה דוגמאות רנדומליות ומנסה למקסם את המנה בין דמיון של זוג הקרוב לסכום הדמיונות בין לבין דוגמאות רנדומליות.

תקציר מאמר:

בשביל להבין את רעיון המאמר במלואו אנו צריכים להכניס עוד מושג חשוב, "עדין משימה" (task refinement).

יעידון משימה: הגדרה ריגורוזית של מושג זה נלקחת מתורת הסיבתיות, אבל לצורך פשטות אפשר זאת ע"י דוגמא. משימת סיווג Y בין זנים שונים של כלבים (או זנים שונים של חתולים) הינה יעידון של משימת סיווג בין כלבים לחתולים Z . כמובן, אם ייצוג הדאטה מספיק טוב בשבל לבצע את Z , הוא יכול מספיק מידע גם בשבל לבצע את Z בצורה טובה.

ולמה בעצם כל זה חשוב, אתם שואלים? קודם כל, נשים לב כי משימת הבדיקה (דיסקרימינציה) בין תכנים שונים בתמונות, כמו שנעשה בשיטות המבוססות NCE, הינה המשימה "הכי מעודנת" עבור דאטה סט נתון. זו הסיבה הנוסףת (קיימים הסברים המצביעים שיטה זו למסום מידע הדדי בין ייצוג דאטה ודאטה עצמו) לכך שהייצוגים שנלמדו בדרך זו, הוכחו שימושיים למשימות downstream שונות. בעצם המאמר מוכיח טענה שלפיה ייצוג אינוריאנטי תחת שינויי סגנון עבור משימה Y נוטר אינוריאנטי לכל משימה Z ש- Y הינה העידון שלה. כאמור, אם האצלחנו ללמידה ייצוג המסוגן לבצע דיסקרימינציה בין תכנים שונים ללא קשר לסגנון, ייצוג זה יעבד טוב גם במקרים downstream שימושם מושוואת על תוכן.

בעצם הוספת איבר רגולרייזציה ללסן הרגיל של InfoNCE תורם להעצמה אי התלות של ייצוגי התמונה בסגנון שלה. כאמור תכונות קרובות ישארו קרובות גם לאחר שינוי סגנון ותכונות רוחקות ישארו כאלו אחריו שינוי סגנון גם כן.

עכשו בואו נבין את המבנה של איבר הרגולרייזציה:

איבר רגולרייזציה - אופן חישוב:

- בונים שני סטיים של פעולות אוגמנטציה (שינויי סגנון) A_1 ו- A_2 , כאשר כל קבוצה מורכבת מזוגות של פעולות אוגמנטציה שונות (a_{1i}, a_{2i}).

לכל דוגמא x :

- עבור כל זוג שינויי סגנון $m-A_i$, משערכים את התפלגות הדמיונות בין ייצוגים שלא תחת a_i ושאר הדוגמאות ממיני-באט' תחת a_i . בשביל זה מפעילים את a_i על x ומחשבים וקטור דמיונות d שלו עם הייצוגים של שאר הדוגמאות תחת a_i . הדמיון מחושב באמצעות מכפלה פנימית של הייצוגים אחרים שchnihim מועברים דרך רשת נירונית רדודה בעלת שכבה אחת או שתיים.
- וקטור d מנורמל כדי להפכו למידת הסתברות המסומנת k_1 .
- מחשבים את וקטורי הדמיונות עבור אוגמנטציות $m-A_2$ באותו צורה: k_2 .
- מחשבים מרחק KL בין k_1 ו- k_2 (דרך מינימית להחליפ' את KL במרקח בין מידות הסתברות ולבדק איך השתנו הייצוגים) ווכ�ים אותו עבור כל זוגות הדוגמאות $m-A_1$ ו- A_2 .

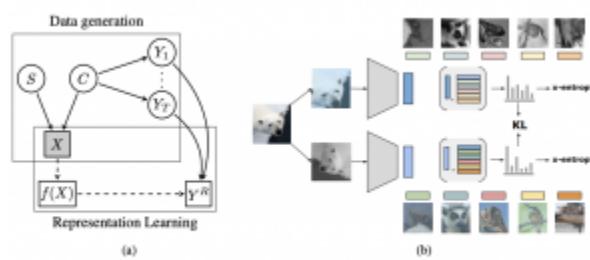


Figure 1: (a) Causal graph formalizing assumptions about content and style of the data and the relationship between targets and proxy tasks. (b) RELIC objective. KL refers to the Kullback-Leibler divergence, while x-entropy denotes cross entropy.

הישגי מאמר:

המאמר הוכיח שיצוגים של RELIC יותר חזקים מalto של שיטות למדידת ייצוג (BYOL, AMDIM, SimCLR) ב-3 היבטים שונים:

1. יחס דיסקרמינטיבי לינארי של פישר (LDR - linear discriminant ratio) המודד מרחק בין היצוגים של הקלאים השוניים. ככל שהמראקים בין מרכזי הקלאים מוכרים של ייצוגים בין הקלאים השונים רוחקים יותר והדיאמטרים של הקלאים קטנים יותר, נקבל LDR גבוה. גובה LDR גבוה של ייצוג הדאטסהט מצביע על קר שנית להבחין בין דוגמאות מהקטגוריות השונות ביותר קלות עי"י מסווג לינארי (ייצוג חזק יותר).

2. ביצועים על משימות downstream שונות (סיווג).

3. זה חדש ומגניב: בחנו את עצמת היצוג על מושית למדידת באמצעות חיזוקים (reinforcement learning) וראו ש- RELIC מצליח לשפר את הביצועים.

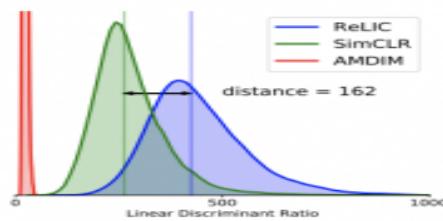


Figure 2: Distribution of the linear discriminant ratio (F_{LDA} , see text) of f for RELIC, SimCLR and AMDIM (y-axis clipped to aid visualization).

Table 1: Accuracy (in %) under linear evaluation on ImageNet for different self-supervised representation learning methods. Methods with * use SimCLR augmentations. Methods with † use custom, stronger augmentations.

Method	Top-1	Top-5
<i>ResNet-50 architecture</i>		
PIRL (Misra & Maaten, 2020)	63.6	-
CPC v2 (Hénaff et al., 2019)	63.8	85.3
CMC (Tian et al., 2019)	66.2	87.0
SimCLR (Chen et al., 2020a)	*	69.3
SwAV (Caron et al., 2020)	*	70.1
RELIC (ours)	*	70.3
InfoMin Aug. (Tian et al., 2020)	†	73.0
SwAV (Caron et al., 2020)	†	75.3
<i>ResNet-50 with target network</i>		
MoCo v2 (Chen et al., 2020b)	71.1	-
BYOL (Grill et al., 2020)	*	74.3
RELIC (ours)	*	74.8

המאמר מציע רעיון מעניין לשיפור ביצועים של שיטות ללמידה הייצוג, המבוססות NCE. הם מציעים להוסיף איבר רגולרייזציה לפונקציית loss הסטנדרטית של NCE. מטרתו של איבר זה היא לגרום לחסמים בין יצוגי תמנונות להיות אינוריאנטיים לשינוי סגנון בתמונות. המאמר מראה שהשיטה המוצעת מצליחה לבנות יצוגים יותר טובים חזקים מאשר שיטות ATA. הינו רוצה לראות שיטה זו מוכלת גם לדומינניים אחרים וגם לסוגים שונים של שימוש.

Review 43: Sharpness-Aware Minimization for Efficiently Improving Generalization

פינת הסוקר:

המלצת קריאה ממייק: חובה לאלו שמתעניינים מה קורה מאחורי הקלעים בתהילך אימון של רשתות ניירוניים.

בהירות כתיבה: גבוהה מאוד.

רמת היכרות עם כלים מתמטיים וטכניקות של DL/ML הנדרשים להבנת מאמר: היכרות טובה עם שיטות אופטימיזציה עבור בעיות עם משתנים רבים.

ישומים פרקטיים אפשריים: שיפור יכולת הכללה של רשתות על ידי החלפת בעית מזעורLoss הרגילה ב-SAM.

פרטי מאמר:

lienek למאמר: [זמן להורדה](#).

lienek לקוד: [כאן](#).

פורסם בתאריך: 04.12.20, בארכיב.

הוזג בכנס: ICLR 2021.

תחום מאמר:

- חקר שיטות אופטימיזציה לאימון של רשתות ניירוניים.

כלים מתמטיים, מושגים וסימונים:

- יכולת הכללה של רשתות ניירוניים.

- Gradient Descent -GD
 - הסיאן (Hessian) של פונקציה.
 - בעיית הנורמה הדואלית (dual norm problem).
-

תמצית מאמר:



Figure 1: (left) Error rate reduction obtained by switching to SAM. Each point is a different dataset / model / data augmentation. (middle) A sharp minimum to which a ResNet trained with SGD converged. (right) A wide minimum to which the same ResNet trained with SAM converged.

המאמר הנזכר מציע ניסוח חדש לבעיית האופטימיזציה המתארחת בזמן אימון רשתות נירוניים. במקום מציאת וקטור משקלים, המציג פונקציית לוס (לסט דוגמאות נתון), המאמר מציע לפחות בעיית אופטימיזציה, שמטרתה למצוא **מינימום סביבתי של פונקציית לוס**. לעומת, במקרים להשתמש בGD לאיתור המינימום המוחלט, ולעדכן את המשקלים לכיוון מינימום אבסולוטי, האלגוריתם המוצע מכוון לנקודה **שבسبיבותה פונקציית הלוס תקבל ערכים מינימליים**.

בנוסף המאמר מוכיח באופן ריגורוזי כי הפתרון בעיית אופטימיזציה שהם מציעים (הנקרא **sharpness aware minimization**) תורם באופן חיובי ליכולת הכללה של המודל המאמן.

רעיון בסיסי:

כמו שאותם בטח יודעים הרוב המוחלט של רשתות הנוירונים המודרניות הין overparameterized בצורה משמעוותית. משתמש מכך כי אופטימיזציה של משקלים רשות על סמרק ערך של פונקציית לוס בנקודה בלבד (!!) עלול להוביל למודלים בעלי יכולת הכללה נמוכה(קרי overfitting). הטענה המרכזית לכך הינה מבנה גיאומטרי מאד מורכב ולא קמור של משטח הלוס. הדוגמא הקלאסית לכך הינה המקירה שבו המינימום של פונקציית לוס "חד" מאד. לעומת אףלו בסביבתה המאוד קרובה של נקודת המינימום הערכיהם של פונקציית הלוס הינם גבוהים משמעותית מערוכה בנקודות המינימום. נקודה מינימום זו עלולה להיות תוצאה של>Dataה רועש ותוביל למודל עם יכולת הכללה נמוכה (overfitting). המאמר מציע פתרון למצב זה עי"י ניסוח בעיית אופטימיזציה שמתמחשת לא רק בערך של פונקציית לוס בנקודה, אלא לוקחת בחשבון את ערכי הלוס בסביבתה. לעומת הניסוח המוצע (SAM) לוקח בחשבון גם את התכונות הגיאומטריות של משטח הלוס בסביבות הנקודה באופן מפורש.

תקציר מאמר:

קיימות מספר רב של שיטות המנסות להגדיל את יכולת הכללה של מודלים בלמידת מכונה. את הפתרונות שהוצעו אפשר לחלק לשתי משפחות עיקריות: הראשונה הינה שינוי האופטימיזיר (Momentum, RmsProp, ADAM ועודמה) והשנייה כוללת שניים בתהילר האימון עצמו (עכירה מוקדמת, BatchNorm, עומק סטטיסטי, אוגמנטציות של DATA ורבה אחרים). שיטות אלו מנסות לפתור את אותה בעית אופטימיזציה של מזעור פונקציית לוס בדרכים שונות. לעומת המאמר הנזכר מציע להחליף את בעית אופטימיזציה עצמה (!!!!).

פרטים טכניים:

פונקציית הלוס המוצעת L מכילה שני איברים - הראשון הוא הלוס המקסימלי בסביבה קטנה של נקודה w (גודלה של סביבה זו הינו היפר-פרמטר) והשני הינו איבר רגולרייזציה סטנדרטי עם נורמת $\|\cdot\|_2$ של w (זה דומה לשיטת אופטימיזציה הנקרואט point proximal point). מעניין כי עבור וקטור משקלים w , ניתן לרשום את $\|\cdot\|_2$ כסכום של הפרש בין הערך המקסימלי של פונקציית לוס בסביבת w (במאמר, הפרש זה נקרא "sharpness" - **חדות**) ואיבר רגולרייזציה חדש שהוא הסכום של נורמת $\|\cdot\|_2$ של וקטור המשקלים w וערך הלוס בנקודה w .

```
Input: Training set  $\mathcal{S} \triangleq \{(x_i, y_i)\}$ , Loss function
 $L: \mathcal{W} \times \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}_+$ , Batch size  $b$ , Step size
 $\eta > 0$ , Neighborhood size  $\rho > 0$ .
Output: Model trained with SAM
Initialize: weights  $w_0$ ,  $t = 0$ ;
while not converged do
    Sample batch  $B = \{(x_j, y_j), \dots, (x_k, y_k)\}$ ;
    Compute gradient  $\nabla_w L_B(w)$  of the batch's training
    loss;
    Compute  $\hat{e}(w)$  per equation 2;
    Compute gradient approximation for the SAM
    objective (equation 3):  $\hat{g} = \nabla_w L_B(w)|_{w+e(w)}$ ;
    Update weights:  $w_{t+1} = w_t - \eta \hat{g}$ ;
     $t = t + 1$ ;
end
return  $w_t$ 
Algorithm 1: SAM algorithm
```

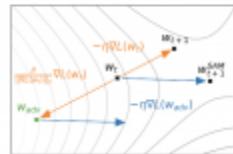


Figure 2: Schematic of the SAM parameter update.

ההיבט התיאורטי:

המאמר הנסקר מוכיח כי עבור סט אימון נתון, הלוס של SAM בכל נקודה w מהווה חסם עליון על הלוס על ה-
population (שממנה סט האימון נדגם) בהסתברות גבוהה (המשפט שਮוכח במאמר טיפה יותר כללי ועובד על
משפחיה יותר רחבה של פונקציות רגולרייזציה). כמוון הכל תחת תנאים טכניים על התפלגות שממנה הדאטאסט
נדגם. בעוד המשפט הזה אומר שפטורן בעית SAM מוביל למודל בעל יכולת הכללה טובה. ההוכחה היא די לא
טריוויאלית ומעורבתחסמי PAC ביסיאנים (מולדים).

פתרונות בעית SAM:

קודם כל משתמשים בקירוב טילור מסדר ראשון, בשביל למצוא את נקודה בסביבה של w עבור הלוס הוא
מקסימלי. אחר כך, הבעה בנידון מתורגמת לעית הנורמה הדואלית הקלאסית, שיש לה פתרון מפורש w_{e} .
אחרי למציבים את w_{e} בביטוי של SAM, מתקבלים בעית אופטימיזציה רגילה (בעית מצעור עם פונקציית מחיר
($w_{\text{e}}(L)$) שפוטרים אותה בדרך הסטנדרטית עם gradient descent. מכיוון w_{e} מכיל את הגרדיינט של
הפונקציה הלוס המקורית L , הביטוי עבור הגרדיינט של ($w_{\text{e}}(L)$) מכיל מטריצת הסיאן (hessian) של L . חישוב
של הסיאן כאשר $L = w$ יש מאות מיליאונים רכיבים זו מושימה מאוד כבדה מבחינה משאבי חישוב זיכרון. אבל
לשםchnerנו, בביטוי מופיעה מכפלה של הסיאן בוקטור, שימושה מאפשר לחשב את הערך של הגרדיינט של
 $(w_{\text{e}}(L))$ ללא חישוב הסיאן. בנוסף לדבר, ניתן להריץ את האלגוריתמים שלהם בדומה ל-GD עם כל גזירה
אוטומטית כמו TensorFlow או PyTorch.

הישגי מאמר:

המאמר הצליח להראות כי הגישה המוצעת מציגה ביצועים עדיפים על פני שיטות אופטימיזציה שונות ומגוונות
(כמו סוגים שונים אוגמנטציה, אופטימיזרים שונים ועוד) על מגוון מאוד רחב של נתונים וארכיטקטורות רשת
שונות. בכל השוואה הם פשוט החליפו את האופטימיזציה המקורית ב-SAM והשו את הביצועים על הטסט סט.

בנוסף, המאמר השווה את ביצועי SAM עבור DATASETS עם ליבלים רועשים וגם אבחן את השינוי בערכים העצמיים של מטריצת הסיאן עבור הפתרון של בעית SAM.

Model	Augmentation	CIFAR-10		CIFAR-100	
		SAM	SGD	SAM	SGD
WRN-28-10 (200 epochs)	Basic	2.7_{±0.1}	3.5 _{±0.1}	16.5_{±0.2}	18.8 _{±0.2}
WRN-28-10 (200 epochs)	Cutout	2.3_{±0.1}	2.6 _{±0.1}	14.9_{±0.2}	16.9 _{±0.1}
WRN-28-10 (200 epochs)	AA	2.1_{±0.1}	2.3 _{±0.1}	13.6_{±0.2}	15.8 _{±0.2}
WRN-28-10 (1800 epochs)	Basic	2.4_{±0.1}	3.5 _{±0.1}	16.3_{±0.2}	19.1 _{±0.1}
WRN-28-10 (1800 epochs)	Cutout	2.3_{±0.1}	2.7 _{±0.1}	14.0_{±0.1}	17.4 _{±0.1}
WRN-28-10 (1800 epochs)	AA	1.6_{±0.1}	2.2 _{±0.1}	12.8_{±0.2}	16.1 _{±0.2}
Shake-Shake (26 2x96d)	Basic	2.3_{±0.1}	2.7 _{±0.1}	15.1_{±0.1}	17.0 _{±0.1}
Shake-Shake (26 2x96d)	Cutout	2.0_{±0.1}	2.3 _{±0.1}	14.2_{±0.2}	15.7 _{±0.2}
Shake-Shake (26 2x96d)	AA	1.6_{±0.1}	1.9 _{±0.1}	12.8_{±0.1}	14.1 _{±0.2}
PyramidNet	Basic	2.7_{±0.1}	4.0 _{±0.1}	14.6_{±0.4}	19.7 _{±0.8}
PyramidNet	Cutout	1.9_{±0.1}	2.5 _{±0.1}	12.6_{±0.2}	16.4 _{±0.1}
PyramidNet	AA	1.6_{±0.1}	1.9 _{±0.1}	11.6_{±0.1}	14.6 _{±0.1}
PyramidNet+ShakeDeep	Basic	2.1_{±0.1}	2.5 _{±0.1}	13.3_{±0.2}	14.5 _{±0.1}
PyramidNet+ShakeDeep	Cutout	1.6_{±0.1}	1.9 _{±0.1}	13.3_{±0.1}	11.8 _{±0.2}
PyramidNet+ShakeDeep	AA	1.4_{±0.1}	1.6 _{±0.1}	10.3_{±0.1}	10.6 _{±0.1}

Table 1: Results for SAM on state-of-the-art models on CIFAR-{10, 100} (WRN = WideResNet; AA = AutoAugment; SGD is the standard non-SAM procedure used to train these models).

ליבלים רועשים:

SAM הציג שיפור ניכר כאשר הוא מופעל באימון על DATASETS עם ליבלים רועשים. בעצם זה לא מפתיע, כי החזק העיקרי של האלגוריתם הוא מניעת התכנסות למינימום "חד", ונוכחות ליבלים רועשים בכמות ניכרת עלול להוביל בקלות למינימום כאלו באלגוריתמים אופטימייזציה קלאסיים.

מבנה ההסיאן בסביבת נקודת אופטימום:

בשביל לאשש את ההנחהות לגבי היכולות של SAM במניעת המינימומים החדים, המאמר בוחן את הערכים העצמיים ("ע"ם המקסימלי ובנוסף גם היחס בין "ע"ם המקסימלי לבין כמה "ע"ם הגבוהים ביותר חוץ מהמקסימלי) של ההסיאן בנקודות אופטימום שנמצאו "ע"ם SAM מול אלו שנמצאו באמצעות אלגוריתמים אחרים. הרוי ידוע שככל שהמינימום יותר חד, יש להסיאן גם ערכים עצמיים גבוהים יותר וגם היחס בין "ע"ם המקסימלי לבין "ע"ם הגבוהים ביותר יותר חד, יש להסיאן גם גובה יותר גם היחס בין "ע"ם המקסימלי לבין ממדים אלו בצורה מאוד משמעותית.

DATASETS:

CIFAR10, CIFAR100, Flowers, Stanford_cars, Birdsnap, Food101, Oxford_IIIT_Pets, FGVC_Aircraft, Fashion-MNIST וכמה אחרים.

ארכיטקטורות רשת שנבחנו:

Wide-ResNet-28-10, Shake-Shake , EffNet, TBMSL-Net, Gpipe

.ג.ב.

מאמר מאד חשוב המציע שיטה מאד מעניינת לשיפור יכולת הכללה של רשותות. לדעתי, יש לשיטה פוטנציאלי רציני להיכנס לארגון כלים סטנדרטי לאימון רשותות. התרשםתי גם המשוואות הרבות והמגוונות מול שיטות אחרות שנעשו במאמר.

Review 44: TransGAN: Two Transformers Can Make One Strong GAN

פינט הסוקר:

המלצת קריאה ממייק: חובה בהחלט (בכל זאת גאון ראשון מבוסס על טרנספורמרים).

בהירות כתיבה: גבוהה.

רמת היכרות עם כלים מתמטיים וטכניקות של ML/DL הנדרשים להבנת מאמר: הבנה טובה בטרנספורמרים וידע בסיסי בGANs.

ישומים פרקטיים אפשריים: TransGAN יודע לייצר תמונות כמו כל גאון אך בינתיים התוצאות אינן נראות בקינה מידה של AOTs בתחום כמו StyleGAN2.

פרטי מאמר:

לינק למאמר: [זמן להורדה](#).

לינק לקוד: [זמן כאן](#).

פורסם בתאריך: 16.02.21, בארכיב.

הציג בכנס: טרם ידוע

תחומי מאמר:

- טרנספורמרים (Transformers)
- GANs (GANs)

כליים מתמטיים, מושגים וסימונים:

- טרנספורמר לתמונות (visual transformers).
- שיטות אוגמנטציה גזירות (differentiable augmentations).
- הוספה של משימה self-supervised (סופר-רצלזיה) לתהליכי אימון.
- אתחול לוקאלי של משקלות self-attention.
- (Frechet Inception Distance (FID

תמצית מאמר:

כפי שאותם בטח יודעים, ב-3 השנים האחרונות הטרנספורמרים השתלטו על עולם ה-NLP. בעקבות המאמר המפורסם "Attention is All You Need", רובם המוחלט של מאמרי-hNL^P משתמשים בארכיטקטורה של הטרנספורמר בצורה זו או אחרת. בשנה האחרונה הטרנספורמרים החלו את פולישתם גם בתחום הראיה הממוחשבת (לדוגמה image is worth 16×16 words). המאמר שסקרתי לאחרונה Transformer (הטרנספורמים הצלחו להפיק ייצוגים representations) חזקית לתחומי המשמשים לאחר מכן למגוון מישימות דיסקרימינטיביות.

המאמר הנזכר מנסה להמשיך向前 לקדם את מהפכת הטרנספורמרים לדומיין הייזואלי ומציג מודל גנרטיבי ראשון שהארכיטקטורה שלו מורכבת מהתרנספורמרים בלבד – ללא שימוש בקונבולוציות. בניית מודל גנרטיבי טוב בדומיין התמונה ללא קונבולוציות זה אכן דבר די מהפכני. הרעיון הקונבולוציות מהוות כלי אולטימטיבי להפקת פיצ'רים מהתמונות, המנצלות את התלות הילוקאלית החזקה שקיים בתמונה אינהרנטי בתמונות. המאמר מציין להסתדר בלבד, וזה אכן בשורה גדולה, אולם יש כאן קטטי קטן. המחברים מצהירים באופן מפורש שהארכיטקטורה שלהם "נטולת קונבולוציות" (CNN-free), ואכן אתם לא תמצאו שם שכבות קונבנציונליות. אבל, וזה אבל די גדול, לקרהת סוף הסקירה אסביר איך הם בכלל זאת הצלחו להכין "חיה מאוד דומה ל-CNN" בדلت האחוריות של המודל שלהם.

הסבר של רעיונות בסיסיים:

המאמר מציע מודל של גאן (GAN) לייצור של תמונות שהגנרטור והדיסקרימינטור שלו מבוססים על הטרנספורמרים.

קצת רקע על גאנים: כפי שאותם זוכרים, גאן מורכב מרשת הגנרטור G, שמטרתה ליצור תמונות ורשת הדיסקרימינטור D, שמטרתה לבדוק בין תמונות אמיתיות לבין אלו שנוצרו ע"י הגנרטור G (מבצע משימת סיווג בינארית). G מנסה לבלב את הדיסקרימינטור ולגרום לו לסווג את התמונות שהוא יוצר כאמתיות. במקרה אחר, הגנרטור מנסה לשפר את איכות הדוגמאות שהוא יוצר על סמך הציון שהוא מקבל מהדיסקרימינטור D.

כאמור, המאמר הנזכר מציע להיפטר מkonvolוציות שהתרגלנו לראותן בgan G והן בדיסקרימינטור D (konvolוציות משוחלפות – transposed convolutions). במקום זאת, המאמר מציע לבנות את G ואת D מטרנספורמרים וגם להוסיף רויבד נספף לתהליכי האימון של הגאן שלהם, שקיביל באופן לא מפתיע את השם TransGAN.

קודם כל, בואו נבין איך ניתן לגנרט תמונה באמצעות הטרנספורמר.

מבנה הגנרטור:

הקלט לגנרטור הינו וקטור ריש גאוסי Z כמו שמקובל גם בגאנים הסטנדרטיים. לאחר מכן התמונה נבנית באופן הבא:

- מעבירים את Z דרך MLP בשביל לבנות את התמונה בrzolozia נמוכה (8×8) כאשר כל פיקסל מיוצג ע"י כמות גדולה של ערכים, המסומנת כ- C.
- לוקחים את הוקטורים המתאים לכל ערך ומוכנסים אותם למקודד (encoder) של טרנספורמר (כל וקטור CAN מיוצג פיצרים של פיקסלים בתמונה שתווצר בהמשך) ביחד עם הקידוד המיקומי הנלמד (learnable positional encoding).
- במשימות NLP, כאשר שם אנו מזינים לטרנספורמר וקטורים המייצגים מילים (או תתי-מילים).
- מבצעים 2×2 sampling sampling באמצעות שיטת pixelshuffle. כתוצאה לכך מתקבלים וקטורים המייצגים פאצ'ים של תמונה בגודל 16×16 , עם מחצית הערכים המקוריים / C.
- חוזרים על שני השלבים האחרונים ומתקבלים כתוצאה לכך וקטורים של פאצ'ים עבור תמונה בגודל של 32×32 , עם $C/4$ ערכים.
- מפעילים הטלה לינארית על הערכים הנוצרים בשביל לבנות תמונה בגודל 32×32 .

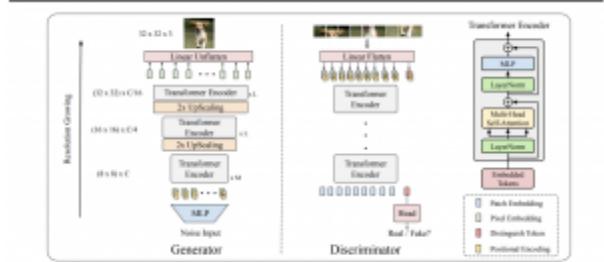


Figure 3. The pipeline of the pure transform-based generator and discriminator of TransGAN. Here $R = W = 8$ and $H_T = W_T = 32$. We show 9 patches for discriminator as an example while in practice we use 8×8 patches across all datasets.

מבנה הדיסקרימינטור:

מכיוון שהדיסקרימינטור צריך לבצע היבין בין תמונה סינטטית (המיוצרת על ידי גנרטור) לתמונה אמיתי, מספיק לקחת פאצ'ים של תמונה ולהכניס אותם למקודד של הטרנספורמר (עם משקלים משלו כמובן). וקטור יציג של פאצ'ים מחושבים באמצעות טרנספורמציה לינארית של הפיקסלים של פאצ'ים. לאחר מכן לוקטור יציג אלו מנוסיף קידוד מיקומי נלמד, והם מוכנסים במספר מקודדים של טרנספורמר אחד אחריו השני. נציין שבדומהה ל-w₂-words An image is worth 16×16 words, מושגים לוקטורי יציג טוקן [cls], שימושו בסופו של דבר לשילוג של תמונה.

איך באמנים TransGAN?

כאן המחברים עשו משהו מעניין. נראה שבהתחלתם הם ניסו לאמן את TransGAN כמו שמאמנים גאנים רגילים אבל התוצאות היו מאכזבות (ניחוש של'). בניסויו להבין את מקור הביצועים החלשים הם החליפו את הגנרטור והדיסקרימינטור של TransGAN (לטירוגון) באלו המבוססים על רשותות קונבולוציה (מ-WGAN-GP, AutoGAN-_{w₂}-StyleGAN), וגילו שמקור החולשה נמצא דווקא בדיסקרימינטור שלא מצילח "לונוט" את הגנרטור שייצור תמונות איכותיות. עקב כך המחברים הושיבו כמה אלמנטים לתהיליך האימון של TransGAN שבפועל שיפרו את ביצועיו בצורה ניכרת.

תוספות לתהיליך האימון:

שימוש בטכניקות אוגמננטציה כבדות: זאת, על מנת ליצור כמות מאוד גבוהה של דוגמאות. הסיבה לכך נראה טמונה בעובדה שלאחר הסרת שכבות הקונבולוציה, bias, human-designed bias, המנצל את התכונות האינגרנטיות של דומיין התמונות, TransGAN לא מצילח למדוד את התכונות האלה בצוותה מספיק.

Table 2. The effectiveness of Data Augmentation (DA) on both CNN-based GANs and TransGAN. We used the full CIFAR-10 training set and DiffAug (Zhao et al., 2020b).

METHODS	DA	IS \uparrow	FID \downarrow
WGAN-GP (GULRAJANI ET AL., 2017)	\times	6.49 \pm 0.09	39.68
	\checkmark	6.29 \pm 0.10	37.14
AUTOGAN (GONG ET AL., 2019)	\times	8.55 \pm 0.12	12.42
	\checkmark	8.60 \pm 0.10	12.72
STYLEGAN v2 (ZHAO ET AL., 2020B)	\times	9.18	11.07
	\checkmark	9.40	9.89
TRANSGAN	\times	6.95 \pm 0.13	41.41
	\checkmark	8.15 \pm 0.14	19.85

אימון משותף של TransGAN עם משימה self-supervised: בנוסף לאימון הרגיל של גן, המחברים הציעו לאמן אותו למשימה של סופר-ריזולוציה. כלומר מורדים את הריזולוציה של תמונה (downsampling) מסט האימון ומנסים לשחזר את התמונה המקורית תוך כדי הוספה של LOSS השחזור (MSE) לLOSS הרגיל של גן.

Table 1. Inception Score (IS) and FID results on CIFAR-10. The first row shows the AutoGAN results (Gong et al., 2019); the second and thirds row show the mixed transformer-CNN results; and the last row shows the pure-transformer GAN results.

GENERATOR	DISCRIMINATOR	IS \uparrow	FID \downarrow
AUTOGAN	AUTOGAN	8.55 \pm 0.12	12.42
TRANSFORMER	AUTOGAN	8.59 \pm 0.10	13.23
AUTOGAN	TRANSFORMER	6.17 \pm 0.12	49.83
TRANSFORMER	TRANSFORMER	6.95 \pm 0.13	41.41

אתחול לוקאלי של משקל self-attention: אטם זוכרים שאמרתי לכם של מרווחים שלא תמצאו שכבות קונבולוציה ב-TransGAN, אך אין הוכנסו פנימה בדلت האחורית? תיכף אסביר זאת. שתי התוספות הראשונות לאימון (סעיפים 1 ו-2) הצלחו לשפר את הביצועים של TransGAN אך הוא עדין נשאר מאחור שיטות SOTA FID ו-IS. עקב כך המחברים הציעו לאותחל משקלים של מנגנון self-attention באופן הבא:

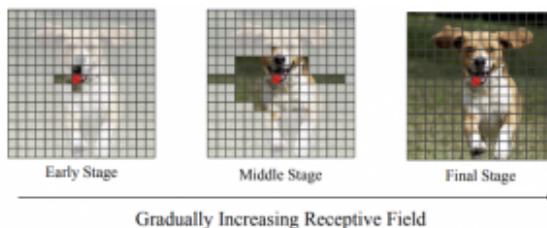


Figure 3. Locality-aware initialization for self-attention. The red block indicates a query location, the transparent blocks are its allowable key locations to interact with, and the gray blocks indicate the masked region. TransGAN gradually increases the allowable region during the training process.

באייטרציות הראשונות מאמנים רק את הזרים הלוקאליים: כלומר מפעילים מסכה של אפסים על מטריצת משקלים של query כר שוקטור יציג של טוקן ("אטץ") "יראה רק את השכנים הקרובים שלו". זה קצת מזכיר את מה שעושים בדקודר של הטרנספורמר הקלסי בשבייל למונע ממנה להתחשב בטוקנים העתידיים בטקסט בעונח. כאן לעומת זאת מונעים מטוקן (= "אטץ") להתחשב בפאצ'ים רחוקים ממנה. ככל שמתקדמים עם אייטרציות האימון מחלישים את המסכות ונותנים יציגי פאצ'ים להתחשב בפאצ'ים רחוקים יותר. לקרהת סוף האימון, מבטלים את המסכות לגמרי ומאמנים את משקל ה-self-attention בזרה רגילה.

תוספת זו למעשה מאפשרת TransGAN להגיע לתוצאות של שיטות SOTA, המוזכרות לעיל.

פינת האינטואיצה לאתחול לokaלי של self-attention:

air שיטת אימון זו קשורה לקונבולוציות אתם שואלים? התשובה פשוטה: כאשר מונעים מהטוקנים ("אטצ'ים") להתחשב בטוקנים רחוקים, אנו למעשה מעניכים ל-TransGAN את מה שנראה bias *human designed bias*: אנו "רמזים" לו שזרים לokaליים מאד חשובים בתמונה. למעשה, אותו bias מוביל אותנו להשתמש ברשותות מבוססות שכבות קונבולוציה כמעט לכל המשימות של הראייה הממוחשבת. כלומר את הקונבולוציות אנחנו לא רואים כאן, אך *human-designed bias* נותר בעינו.

הישgi מאמר:

- AutoGAN, WGAN-GP זמינים דומים של שיטות SOTA חזקות כמו TransGAN ו- StyleGAN v2.

Table 5. Unconditional image generation results on CIFAR-10.

METHODS	IS	FID
WGAN-GP (GULRAJANI ET AL., 2017)	6.49 ± 0.09	39.68
LRGAN (YANG ET AL., 2017)	7.17 ± 0.17	-
DFM (WARDE-FARLEY & BENGIO, 2016)	7.72 ± 0.13	-
SPLITTING GAN (GRINBLAT ET AL., 2017)	7.90 ± 0.09	-
IMPROVING MMD-GAN (WANG ET AL., 2018A)	8.29	16.21
MGAN (HOANG ET AL., 2018)	8.33 ± 0.10	26.7
SN-GAN (MIYATO ET AL., 2018)	8.22 ± 0.05	21.7
PROGRESSIVE-GAN (KARRAS ET AL., 2017)	8.80 ± 0.05	15.52
AUTOGAN (GONG ET AL., 2019)	8.55 ± 0.10	12.42
STYLEGAN V2 (ZHAO ET AL., 2020B)	9.18	11.07
TRANSGAN-XL	8.63 ± 0.16	11.89

Table 1. Inception Score (IS) and FID results on CIFAR-10. The first row shows the AutoGAN results (Gong et al., 2019); the second and thirds row show the mixed transformer-CNN results; and the last row shows the pure-transformer GAN results.

GENERATOR	DISCRIMINATOR	IS↑	FID↓
AUTOGAN	AUTOGAN	8.55± 0.12	12.42
TRANSFORMER	AUTOGAN	8.59± 0.10	13.23
AUTOGAN	TRANSFORMER	6.17± 0.12	49.83
TRANSFORMER	TRANSFORMER	6.95 ± 0.13	41.41

.נ.ב.

מאמר מואוד מעוניין המציע גאן ראשון מבוסס כלו טרנספורמרם המגיע לביוצוי SOTA. תוצאה זו הושגה בזכות שימוש בכמה טרייקים מעוניינים במהלך האימון.

Review 45: Rethinking Attention With Performers

פינת הסוקה:

המלצת קריאה ממיק: חובה לחובבי הטרנספורמרם ולאנשי NLP.

bahiorot כתיבה: גבוהה.

רמת היכרות עם כלים מתמטיים וטכניקות של ML/DL הנדרשים להבנת מאמר: נדרשת היכרות בסיסית עם תורת הקרנלים, הבנה טוביה בפעולת ליבה בטרנספורמרם (self-attention).

ישומיים פרקטיים אפשריים: ניתן להשתמש בגישה המוצעת במאמר עבור כל שימושה בה הסיבות הריבועית של מנגנון self-attention של הטרנספורמר הינה בעיה מבחינה משאבי חישוב.

פרטי מאמר:

lienק למאמר: [זמן להורדה](#).

lienק לקוד: [זמן CAN](#).

פורסם בתאריך: 09.03.21, בארכ'יב.

הוזג בכנס: ICLR 2021.

תחומי מאמר:

- טרנספורמרים בעלי סיבוכיות חישובית נמוכה.

כלים מתמטיים, מושגים וסימונים:

- מנגנון self-attention - SA .
- קרNELי סופטמקס (softmax kernels) (soft).
- פיצ'רים חיוביים אורותוגונליים רנדומליים (Positive Orthogonal Random Features).

מבוא ותמצית מאמר:

טרנספורמר הינו ארכיטקטורה של רשתות נירוניים عمוקות שהוצאה בשלבי 2017 במאמר "Attention is what we need". מאז הטרנספורמים כבשו את עולם ה-NLP והפכו לארכיטקטורה כמעט דיפולית בתחום. רוב המוחלט של מאמרי NLP של הימים האחרונים משתמשים בטרנספורמים בצורה זו או אחרת. לאחרונה הטרנספורמרים התחילו לפול אטדרם גם לדומין הייזיאלי והופיעו בכמה מאמריהם שחקף סקרת' ([Image is Worth 16x16 Words](#), [TransGAN](#), [Image Processing Transformer](#)).

הקלט לטרנספורמר הינו סט או סדרה של עצמים (מילה, תת-מילה, פאץ' בתמונה, דגימות אודיו וכו') שכל אחד מהם מיוצג על ידי וקטור. הלב של הטרנספורמר הינו מנגנון self-attention כימות וקטורים בין איברים שונים בסט ובסדרה. המטרה של הטרנספורמר הינה הפקה של ייצוג וקטורי של כל איבר בסדרה/סט, התלוי באיברים האחרים (מה שנקרא contextualized embedding ב-NLP). דרך אגב לאחרונה יצא [आטח](#), שהראה שהכוח של מנגנון self-attention נבע משלילבו עם skip-connections ושכבות fully-connected. בנוסף נציג כי כאשר הקלט הינו בעל סדר אינהרנטי בין איבורי (כמו טקסט או תמונה), אז מוסיפים וקטור הייצוג של כל איבר גם וקטור המכיל מידע על מיקומו בסדרה (PE - Positional encoding). כאשר הקלט הינו סט ללא חשיבות מסוימת (אינוריאנטי ל-transformer), לא נדרש PE.

מהחר בשלב הראשון מנגנון SA מחשב את הדמיון של כל איבר אחר בסדרה, הסיבוכיות של שלב זה הינה ריבועית במונחי אורך הסדרה (נסמן את אורך הסדרה ב-L). סיבוכיות זו עלולה להיות בעייתית עבור סדרות ארוכות מבחינה משאבי חישוב וזכרון הנדרשים. לעומת זו מחריפה עבור ארכיטקטורות המורכבות ממספר שכבות של טרנספורמרים. אגב, סוגיה זו מהווה את אחד המכשולים העיקריים (בנוסף לכך שהטרנספורמר, בצורתו הקלאסית, לאبني לניצול קשרים לוקאלים הקיימים בתמונות, אך זה ניתן לטיפול על ידי שימוש שיטות אימון מתחכם) המונעים את השתלטות הטרנספורמרים גם על הדומין הייזיאלי. הסיבה לכך טמונה במספר הפאצ'ים (איברים בסדרה) הגובה בתמונה ברוחזוץ גובהה - המימוש הסטנדרטי של מנגנון SA עלול להיות כבד מאוד במקרה חישובית והן מבחינת הזכרן הנדרש).

בשנה האחרונה יצא מספר מאמרים שהציגו וריאנטים זולים יותר חישובית של הטרנספורמר כמו [Lformer](#) ו-[Reformer](#). כדי להוריד את הסיבוכיות הריבועית של הטרנספורמר, רוב המאמרים הינו הנו על תכונות של הקשרים בין האיברי הסדרה או על מטריצות K, Q ו-V המשתתפות בחישוב של SA. לטענת מחברי המאמר הנוסף כל הוריאנטים "קלים חישובית" של הטרנספורמר, שנבדקו על ידייהם, הפגינו ביצועים ירודים ממשמעיתיחסית לגורסתו המקורית (היקפה חישובית) של הטרנספורמר. המאמר טוען שהסיבה לביצועים חלשים אלו הינה אי-קיים של התנאים עליהם מtabseet וריאנטים אלו.

כותבי המאמר אינם מניחים שום הנחה על תכונות/מבנה של הקשרים בין איברים ומציעים מסגרת מתמטית ריגורוזית למציאת קירוב למטריצת attention (המחושבת על ידי מנגנון SA) **בבסיסיות לינארית במנוחי אורך הקלט**. בנוסף, ניתן לשחק עם הפרמטרים של קירוב זה ולהגיע לכל דיק רצוי בשערוך של מטריצת attention. יתרה מזזו, המאמר מוכיח כי שקיירוב זה הינו:

- אומדן בלתי מוטה (או ממש קרוב לזה) למטריצת attention.
- מתכנס בצורה יוניפורמית (אותה מהירות התכנסות לכל איבר) למטריצת attention ולכל טווח ערכי הה-attention).
- בעל שונות נמוכה.

הסבר של רעיונות בסיסיים:

כאמור בשלב הראשון של חישוב מטריצת softmax, פועלות softmax מחושבת על מכפלת המטריצות \mathbf{Q}^* ו- \mathbf{K} (משוחלפת). מטריצות \mathbf{Q}^* ו- \mathbf{K} מורכבות מהמכפלות של מטריצות ידי Query ומטריצת Key (המסומנות על ידי \mathbf{Q} ו- \mathbf{K} בהתאם) על וקטורי הייצוג של הקלט \mathbf{q} ו- \mathbf{z}_k . למעשה כל המכפלות הפנימיות מנורמלות ב- $^{1/2}d$ אך זה לא משנה את עיקרי החישוב. כמובן פועלות softmax מופעלת על המטריצה (נסמן אותה כ- \mathbf{A}), שアイיר [3] שלה הינו מכפלה פנימית של וקטורי \mathbf{q} ו- \mathbf{z}_k . נציין שגודלו מטריצה זו היא $L \times L$, כאשר L הינו אורך הקלט. לאחר מכן, המכפלה הפנימית של מטריצה \mathbf{A} מכפלת במטריצה \mathbf{V} שבונה מכפלות של וקטורי הייצוג האיברים במטריצת \mathbf{V} (מטריצת Value). הגודל של מטריצת \mathbf{V} הינו $D \times L$, כאשר D הינו מימד של וקטורי הייצוג. ניתן לראות כי סיבוכיות זמן וגודלו זכרון הנדרש הם $O(L^2D)$. וזה לב הבעה עם הטרנספורמטורים עבור קלט ארוך כמו פסקה שלמה של טקסט או כל הפקטים של תמונה ברזולוציה גבוהה. המאמר מציע שיטה לקרב את החישוב של softmax על המכפלה של \mathbf{Q}^* ו- \mathbf{K} משוחלפת על ידי מכפלה של שתי מטריצות \mathbf{Q}' ו- \mathbf{K}' בגודל של $L \times L$, כאשר L הרבה יותר קטן מ- L . זה מאפשר להחליף את סדר המכפלה של המטריצות בחישוב SA:

1. מכפילים מטריצה \mathbf{V} בגודל $D \times L$ במטריצה \mathbf{K} משוחלפת בגודל $L \times L$. כתוצאה לכך מקבלים מטריצה ' \mathbf{A}' בגודל $D \times L$.
2. מכפילים את המטריצה ' \mathbf{A}' שנתקבלה במטריצה ' \mathbf{Q}' בגודל $L \times L$.

קל לראות שהסיבוכיות של הזיכרון ושל החישוב במקרה זה אינה לינארית ב- L (כאשר $L > D$).

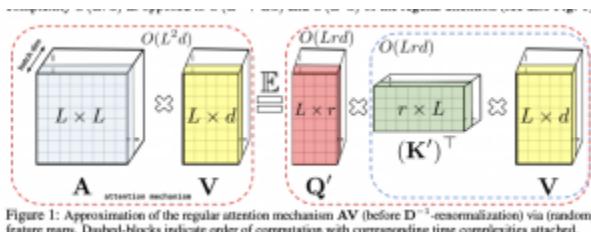


Figure 1: Approximation of the regular attention mechanism AV (before D^{-1} -renormalization) via (random) feature maps. Dashed-blocks indicate order of computation with corresponding time complexities attached.

אבל השאלה המהותית כאן היא: איך ניתן לבנות מטריצות ' \mathbf{Q}' ו- \mathbf{K}' כך שמכפלתן תהווה קירוב בעל תכונות המוזכרות לעיל (בלתי מוטה, בעל קצב התכנסות יוניפורמית שונות קטנה). מחברי המאמר מציעים שיטה, הנקראת FAVOR++, לקירוב של מטריצה \mathbf{A} , שאיבריה הם ערכי softmax ה-softmax המרכיבים שלו הם המכפלות הפנימיות של וקטורי \mathbf{q} ו- \mathbf{z}_k . למעשה, המאמר מציע שיטה יותר כללית לקירוב של כל פונקציה מהצורה $\mathbf{k}, \mathbf{q}, \mathbf{K}$, אשר \mathbf{K} זה קרבן (פונקציית בעלות תכונות מסוימות) חיובי. הקירובamusה מהוות תוחלת של מכפלה פנימית של $(\mathbf{q})\mathbf{f}$ ו- $(\mathbf{k})\mathbf{f}$ (מסומנת \mathbf{E}) כאשר \mathbf{f} הינה פונקציה אקראית (randomized) מ- \mathbb{R}^d ל- \mathbb{R} . ד"א

זה די מזכיר ייצוג קרNEL באמצעות [Random Fourier Features](#) למי שמכיר. המאמר מציע לנקוט את פונקציית מהצורה הבאה:

$$\phi(\mathbf{x}) = \frac{h(\mathbf{x})}{\sqrt{m}}(f_1(\omega_1^\top \mathbf{x}), \dots, f_1(\omega_m^\top \mathbf{x}), \dots, f_l(\omega_1^\top \mathbf{x}), \dots, f_l(\omega_m^\top \mathbf{x})) \quad (1)$$

כאשר

- $\omega_1, \dots, \omega_m$ הינם פונקציות $\mathbb{R} \rightarrow \mathbb{R}$.
- h הינה פונקציה $\mathbb{R}^d \rightarrow \mathbb{R}$.
- $m, l = 1, \dots, d$ - הינם וקטורים, המוגדרים (פעם אחת לאורך כל החישוב) מהתפלגות D על \mathbb{R}^d . ברוב המקרים התפלגות D הינה איזוטרופית, כלומר פונקציית ההתפלגות שלה קבועה על סferה (sphere).

לדוגמא, אם נkeh () f₁=cos(), f₂=sin(), h=1, ו-D הינה התפלגות גאוסית סטנדרטית, אז נקבל קירוב של מה שנזכר [קרNEL גauss](#): $K_{\text{gauss}}(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|^2/2)$ (עד כדי הנרמול). אם נשים לב כי

$$SM(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|^2/2) \quad (2)$$

אז קל להראות כי $SM(\mathbf{x}, \mathbf{y})$ ניתן לקירוב על ידי פונקציה, המוגדרת על הפונקציות הבאות באמצעות הנוסחה (1):

$$h(\mathbf{x}) = \exp(-\|\mathbf{x}\|^2/2), f_1(\mathbf{x}) = \cos(\|\mathbf{x}\|), f_2(\mathbf{x}) = \sin(\|\mathbf{x}\|) \quad (3)$$

אז למעשה הצלחנו לקירב את איברי מטריצות \mathbf{Q}^* ו- \mathbf{K}^* משוחלפת על ידי מכפלה פנימית של וקטורים, המחשבים מוקטורי \mathbf{q}_i ו- \mathbf{v}_j (עם פונקציית $h(\cdot)$). נוכל אז לבצע את מכפלת המטריצות בביטוי של מטריצת attention בסדר אחר, ובכך הורדנו את הסיבוכיות לינארית במנוחה אורך הקלט. אבל יש פה קצת' קטן: softmax למעשה יותר כירוף לינארי קמור (שכל מקדמי חיבויים ומונומלים) של המכפלה של \mathbf{Q}^* ו- \mathbf{K}^* משוחלפת. כאשר אנו מחליפים את החישוב הזה על ידי הקירוב שיכל לקבל כל ערך (גמ' שלילי), זה עשוי להיות בעייתי ולגרום לאי דיקונים רציניים במיוחד במקרים מסוימים. אם נזכיר שsoftmax מודד דמיון בין וקטור \mathbf{q}_i ו- \mathbf{v}_j key בין איברים שונים, אז סביר להניח שרוב ערכי \mathbf{q}_i יהיו קרובים לאפס. המאמר גם מראה שאם משתמשים בקירוב (3) אז אי הדיקונים של הקירוב, יחסית לערכים האמיתיים של softmax, הינם די משמעותיים.

כלומר לא רק שאנו צריכים לקירב את החישוב של softmax אלא לעשות זאת באמצעות פונקציות לא שליליות. המאמר מציע להשתמש בקירוב הבא:

$$SM(\mathbf{x}, \mathbf{y}) = \mathbb{E}_{\omega \sim \mathcal{N}(0, I_d)} \left[\exp \left(\omega^\top \mathbf{x} - \frac{\|\mathbf{x}\|^2}{2} \right) \exp \left(\omega^\top \mathbf{y} - \frac{\|\mathbf{y}\|^2}{2} \right) \right]$$

שניינן על ידי

$$h(\mathbf{x}) = \frac{1}{\sqrt{2}} \exp(-\frac{\|\mathbf{x}\|^2}{2}), l = 2, f_1(u) = \exp(u), f_2(u) = \exp(-u)$$

המאמר מראה שקירוב ה-softmax דרך הביטוי, הנתון על ידי שתי המשוואות האחרונות, מצליח לקירב את הערכים האמיתיים של מטריצת attention בצורה יוניפורמית ועם שונות נמוכה. כדי לגרום לקירוב להיות יותר

מודוק בהינתן אותו מספר של וקטורים המוגבלים מהתפלגות גאוסית סטנדרטית $\mathcal{N}(0, I)$ (פעם אחת בלבד לאורך כל הערך), המאמר מציע לבצע תהליך אורתוגונליזציה של וקטורים אלו. אחד הדרכים לעשות זאת היא להשתמש בשיטת [גרם-שmidt](#).

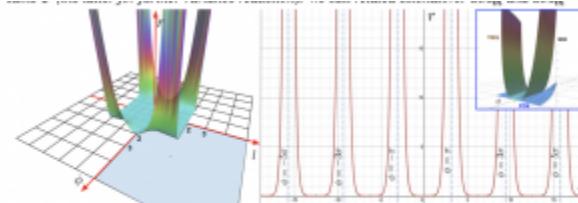


Figure 2: Left: Symmetrized (around origin) utility function r (defined as the ratio of the mean squared errors (MSEs) of estimators built on trigonometric and positive random features) as a function of the angle ϕ (in radians) between input feature vectors and their lengths l . Larger values indicate regions of (ϕ, l) -space with better performance of positive random features. We see that for critical regions with ϕ large enough (small enough softmax-kernel values) our method is arbitrarily more accurate than trigonometric random features. Plot presented for domain $[-\pi, \pi] \times [-2, 2]$. Right: The slice of function r for fixed $l = 1$ and varying angle ϕ . Right Upper Corner: Comparison of the MSEs of both the estimators in a low softmax-kernel value region.

לבסוף, המאמר מוכיח בצורה ריגורוזית (באמצעות כלים די לא טריוויאליים את התכונות התיאורתיות "הטובות" של הקירוב הזה (רוב המאמר זה הוכחות - בערך 30 עמודים).

הישגי מאמר:

המאמר הראשון (למיון ידועתי) שהצליח להקטין את סיבוכיות החישוב (והאsson) של מטריצת softmax בטרנספורמר ליניארית במונחי אורך סדרת הקלט ללא הנחות כלשהן על מטריצות Key, Query, Value ועל ערכי attention עצם.

נ.ב.

מאמר מציע שיטה להקטין את סיבוכיות של הטרנספורמר ליניארית ומוכיח את כל טענותיו גם (!!!) בצורה ריגורוזית. המאמר לא פשוט לקרוא אך לשמה לנו כדי להבין את העיקר לא צריך להתעמק בפרטיה הוחחות (5-6 העמודים הראשונים מספקים).

Review 46: Discriminator Rejection Sampling

פינת הסוקר:

המלצת קריאה ממיליק: חובה לאהובי גאניםandi מומלץ עברו האחרים.

בהירות כתיבה: בינוי פלאס.

רמת היכרות עם כלים מתמטיים וטכניקות של DL/ML הנדרשים להבנת מאמרה: הבנה בשיטות האימון של גאנים, ידע בסיסי בשיטות דגימה כמו [Rejection Sampling](#).

ישומים פרקטיים אפשריים: ג'ינרט תמונות יותר אינטראקטיב עם גאנים.

פרטי מאמר:

lienק למאמר: [זמן להורדה](#).

lienק לקוד: [לא אוטר](#).

פורסם בתאריך: 26.02.19, בארכ'יב.

הוגג בכנס: ICLR 2019.

תחום מאמר:

- חקר שיטות ג'ינרטור דוגמאות באמצעות גאנים מאומנים.

כלים מתמטיים, מושגים וסימונים:

- גאנים (GANs).
- Rejection Sampling

תמצית מאמר:

המאמר מציע שיטה לשיפור איכות התמונה המוגנרטות על ידי GAN מאומן, תוך כדי ניצול "המידע" שנוצר בדיסקרימינטור (D) במהלך תהליכי האימון של ה-GAN. נזכיר ש- D מאומן להבחין בין התמונה המוגנרטות על ידי הגנרטור G לבין התמונה מט האימון. הפלט של D הינו הסתברות שהקלט הינו של תמונה אמיתית (מט האימון). המאמר מציע לנצל את התפלגות על מרחב התמונה המשוררת על ידי D (באופן לא מפורש) בשילוב לתיקן את התפלגות התמונה המשוררת על ידי G (התמונה המוגנרטות) ובכך לשפר את איכותו של התמונה המוגנרטות.

רעיון בסיסי:

כאשר D מאומן טוב מספיק, הוא משרה התפלגות על מרחב התמונה בעל התכונות הבאות:

- **תמונות "שנראות דומות לטבעיות"** מקבלות הסתברויות גבוהות.
- **תמונות שנראות "לא אמיתיות"** מקבלות הסתברויות נמוכות.

ב翦ורו הרבה יותר הגיוני לדגום מההתפלגות המשוררת ע"י D כי אז אנו נדגום תמונות, שנראות דומות לאמיתיות (אלו ש- D מעניק להם הסתברות גבוהה), בסבירות יותר גבוהה. אבל איך נוכל לדגום מההתפלגות הזאת, אם היא לא ניתנת לנו בצורה מפורשת (intractable)? כדי להתגבר על קשי זה, מחברי המאמר **משתמשים בדגימות של G ומפעלים טכניקת דגימה הנקראת (rejection sampling)** בשילוב לדגום מההתפלגות המשוררת על ידי D .

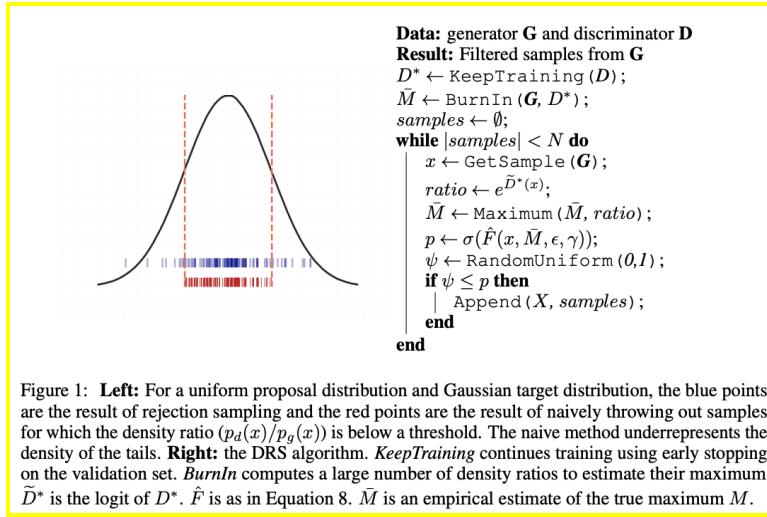


Figure 1: **Left:** For a uniform proposal distribution and Gaussian target distribution, the blue points are the result of rejection sampling and the red points are the result of naively throwing out samples for which the density ratio ($p_d(x)/p_g(x)$) is below a threshold. The naive method underrepresents the density of the tails. **Right:** the DRS algorithm. *KeepTraining* continues training using early stopping on the validation set. *BurnIn* computes a large number of density ratios to estimate their maximum. \tilde{D}^* is the logit of D^* . \hat{F} is as in Equation 8. \bar{M} is an empirical estimate of the true maximum M .

תקציר מאמר:

כאשר מסתכלים על הרעיון הזה בצורה עמוקה יותר, עלות מספר שאלות לגבי יעילותו. הרי אם היה מידע מועיל כלשהו במשקלים של D , הוא היה מועבר ל- G -GAN האימון של הרעיון "AINFORMCIA" מ- D -GAN דרך הגרדיינט של פונקציית הלוס של GAN). אולם יש מספר סיבות למה בפועל לא כל האינפורמציה הנזכרת ב- D מועברת בסופו של דבר ל- G .

1. ההנחה על תהליכי האימון של GAN לא תמיד מתקינות (למשל הן ל- D והן ל- G יש קיולות סופית ולא תמיד ניתן להעביר את כל האינפורמציה מאחד לשני דרך המשקלים של רשותות אלו).
2. יתכן שקיים מצבים בהם יותר קל ל- D להבדיל בין התפלגות נכונה ללא נוכנה (על סמך הדגימות) מאשר לממד התפלגות נכונה ב- D .
3. הסיבה הכי פשוטה: יתכן שאנו לא מאמנים GAN מספיק זמן בשביל ש- G יהיה מסוגל לממד את ההתפלגות האמיתית. ככלומר האימון נגמר לפני של האינפורמציה מ- D מועברת ל- G .

נתחילה עם הסבר קצר על rejection sampling המהווה את אבן היסוד של הרעיון המוצע במאמר:

:Rejection sampling(RS)

טכנית זו מיועדת לדגימה מההתפלגות d , שהדגימה הישירה ממנה קשה (למשל מההתפלגות שניתנה בצורה לא מפורשת). במקום זאת, דוגמים מההתפלגות אחרת g , המוגדרת מעל אותו מרחב, שנייתן לדוגם ממנה אם מתקיים התנאי הבא: המקסימום של החיסכון בין הערכים של d ושל g צריך להיות חסום ע"י קבוע M . אז איך זה בעצם עובד? דוגמים מ- g נקודה y ומחשבים את הערך של $d(y)$ ב- d מוכפל ב- M כלומר מחשבים y את הדגימה y בהסתברות t ודוחים אותה בהסתברות $t - 1$.

נסמן ב- d את ההתפלגות המושרota על ידי D -GAN. כעת נשאלת השאלה איך אנו בעצם נבצע RS אם אנו לא יודעים לחשב לא את d ולא את g בצורה מפורשת? הטריך הוא **שאנו צריכים לחשב את המנה**

ולא את הערכים עצם. המאמר מצין, שתחת תנאים מסוימים ("התנאים האידיאליים") על d ו- d_g , ניתן לדגם את d דרך d_g בצורה מדויקת.

התנאים האידיאליים:

1. d ו- d_g יש אותו סט תומך (כלומר הן שונות מ-0 באותה הנקודות).
2. הקבוע M (המקסימום של היחס בין d ו- d_g) ידוע או ניתן לחשב אותו.
3. $L-G$ נתון, ניתן לאמן את D עד להבאתו לערכו המינימלי האבסולוטי התיאורטי של פונקציית הלוס של גאן (הערך הזה שווה ל-4 \log_2). כמובן שזה בלתי אפשרי כי יש לנו דאטאטים בגודל סופי והאימן שלנו הוא גם באורך סופי.

תחת תנאים אלו המאמר מראה כי ניתן לדגם מ- d דרך d_g באמצעות RS. הנוסחה עבור המנה של d ו- d_g במקרה זה כוללת את האקספוננט של הלוג'יט (logit) של הדיסקרימינטור האופטימלי * D (ה- D האופטימלי מוגדר בתור זה שambil את פונקציית הלוס למינימום האבסולוטי). ההוכחה היא כי אלגנטית ומינצת את הנוסחה עבור הערך האופטימלי של D בנקודה x (השווה ליחס בין $(x_d + p_d)$ לבין $(x_g + p_g)$ עבור G קבוע, המופיע על ידי $(x)^*$).

כמובן שאף אחד מהתנאים אלו לא מתקיים במציאות. המאמר מציע דרך לבצע RS למראות אי קיום התנאים האידיאליים.

לגביה תנאים (1) ו- (3) המאמר טוען כי ניתן להשתמש ב- D מאומן מספיק טוב כקירוב טוב של D . אם מאמנים את D בצורה "המנועת" overfitting (רגוליזציה, עזרה מוקדמת וכדומה - לשון המאמר). במקרה זה D המאומן יודע להבדיל בין דוגמה "טובה" לדוגמה רעה גם אם הדגימות האלו יהיו בעלות הסתברות 0 עבור d_g האופטימלי (עבור D). הם גם מוכחים הנחה זו אמפירית.

לגביה (2) הם מציעים לשערק קבוע M בשני שלבים: שלב השערוך שבו הם מחשבים את הערך של M על K0. דגימות ראשונות (ניתן להראות כי עבור דוגמא נתונה M זה האקספוננט של לוג'יט הערך של D עבור דוגמה זו). אחר כך בשלב הדוגמה הם מעדכנים את הערך של M אם מתקבל ערך גבוה יותר של M עבור אחת הדגימות. זה עשוי להוביל לשערוך יתר של הסתברויות קבלת דגימות שקדמו לעדכן של M אך לטענת המאמר עדכן של M לא קורה באופן תמידי בפועל.

בנוסף המאמר מצין כי ל-RS יש בעיה לדגום מרחבים בעלי מיד גובה כי הסתברות לקבלת דוגמה ϵ היא מאוד קטנה. המחברים מציעים טרייק יפה (שambilן מעוזות "קצת" את ההתפלגות האמיתית של הסתברויות קבלת הדוגמה) כדי "להתגבר" על הבעיה זו. הטרייק הוא להשתמש פרמטריזציה של הביטוי עבור הסתברויות קבלת הדוגמה: מכניםים פרמטר φ האחראי על "הרחבה" סט הערכים של הסתברויות זו. כלומר אם ערך הפרמטר גובה ϵ נוטה לקבל ערכים גבוהים יחסית וכך אשר ערכו של φ נמוך, גם ϵ נוטה להיות נמוך ורוב הדגימות נדחות. בסוף עושים אופטימיזציה על הערך של פרמטר זה.

הישgi מאמר:

המחברים הצליחו לשפר את יכולות התמונהות המגונרטות ע"י GAN עם השיטה שלהם. ההשוואה בוצעה מול SAGAN שהוא SOTA על ייצור תמונות (מאומן על Imagenet) לפני כשנתיים.

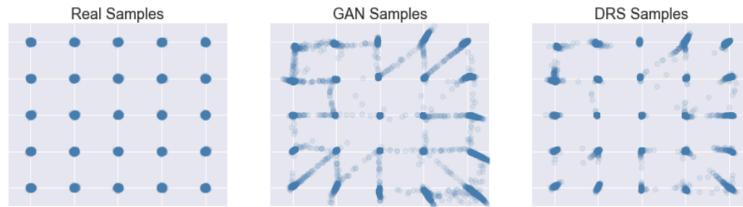


Figure 3: Real samples from 25 2D-Gaussian Distributions (*left*) as well as fake samples generated from a trained GAN model without (*middle*) and with DRS (*right*). Results are computed as an average over five models randomly initialized and trained independently.

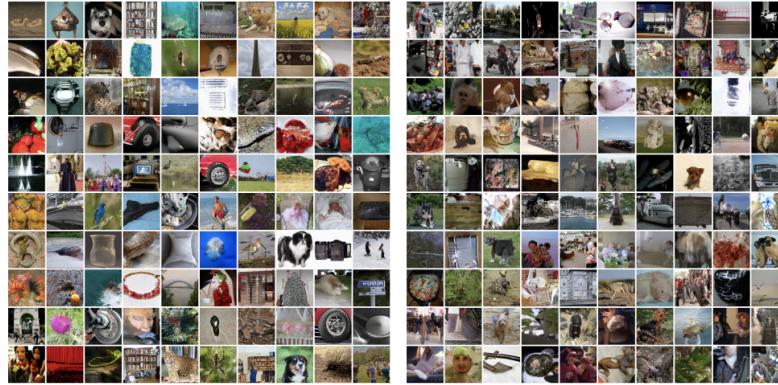


Figure 4: Synthesized images with the highest (*left*) and lowest (*right*) acceptance probability scores.

מטריקות השוואה:

.Frechet Inception Distance, Inception Score

ג.ב.

מאמר עם רעיון מבריק. למרות התוצאות המרשימות, חסרות בו הוכחות ריגורזיות של ההנחות שלהם ואני מקווה שיבואו בהמשך.

שנקרא:

Review 47: Perceiver: General Perception with Iterative Attention

פינת הסוקר:

המלצת קריאה ממיל'ק: חובה (!!)- לאוהבי הטרנספורמרים, לאחרים מומלץ מאד (הרעיון ממש מגניב).

בהירות כתיבה: בינויית פלאס.

רמת היכרות עם כלים מתמטיים וטכניקות של DL/ML הנדרשים להבנת מאמר: היכרות בסיסית עם ארכיטקטורת הטרנספורמר וידע בסיסי בסיבוכיות.

ישומים פרקטיים אפשריים: טרנספורמרים בעלי סיבוכיות נמוכה המותאמים לעיבוד סדרות ארוכות של נתונים (פאטצ'ים של תמונה, פרימים של וידאו, טקסט ארוך וכדומה).

פרטי מאמר:

lienק למאמר: [זמן להורדה](#).

lienק לקוד: [כאן](#), [כאן](#) [כאן](#) (לא רשמיים).

פורסם בתאריך: 24.11.20, בארכיב.

הוגג בכנס: NeurPS

תחום מאמר:

- טרנספורמרים בעלי סיבוביות חישוב ואחסון נמוכות.

כלים מתמטיים, מושגים וסימונים:

- יסודות ארכיטקטורת הטרנספורמרים.

מבוא:

הטרנספורמר הוא ארכיטקטורה של רשתות נוירוניים המיועדת לעיבוד של>Data סדרתי. הטרנספורמרים הוצעו במאמר משנת 2017 הנקרא [Attention is All You Need](#). מאז השתלטו הטרנספורמרים על עולם ה NLP והפכו לארכיטקטורת ברירת המחדל שם. הטרנספורמרים משמשים לבניית "צוגי>Data חזקים (pretraining) שלאחר מכן ניתן לצליל אותם (fine tuning) למגוון שימושות downstream.

בתקופה الأخيرة, התחילו הטרנספורמרים את פליישתם גם בתחום הראייה הממוחשבת. בין המאמרים שהשתמשו בטרנספורמרים למשימות שונות בדומיין התמונה ניתן למנות ([An image is worth 16×16](#)) (Pretrained Image Transformer) ו- ([DETR](#), [TransGAN](#), [words](#)). שלושה מאמרים שスクרנו לאחורינה ([Knowledge Vision Transformers](#)) לאחרונה אנחנו רואים שימוש בטרנספורמרים גם למשימות עיבוד וידאו נזכיר שבדרך כל הקלט לטרנספורמרים במשימות הראייה הממוחשבת הינט הפעתיים של תמונה הקולט.

עם זאת קיימים מספר אתגרים המונעים שימוש נרחב יותר בטרנספורמרים בדומיין הייזואלי.

• התליות הלוקאליות האינגרנטיות שקיימות בתמונות.

רשתות קובולוציה, "המככבות" כמעט בכל משימה של הראייה הממוחשבת, מנצלות את התליות (קשרים) הלוקאליות הקיימות בתמונות על ידי שימוש בפיקסלים סמוכים בלבד לחישוב פיצ'רים בשכבות הנמוכות. לעומת זאת, בניית הטרנספורמרים אינו מאפשר לבנות "צוגים לokaליים" מאחר וייצוג הדטה בטרנספורמר הקלאסי נבנה באמצעות **ניתוח קשרים בין כל חלק הדטה בו זמן** (להסביר מפורט על הטרנספורמר ראו [TransGAN](#)). על Koshi זה ניתן להתגבר על ידי מגנון אתחול משקלים

מתוחכם (ראה [TransGAN](#)). יש לבדוק שמשתמשות בשכבות קונבולוציה כשלב מקדים לבניית "יצוגים של פאטיים לפני החתמת לטרנספורמר").

- **סיבוכיות חישובית ריבועית של הטרנספורמרם במונחי אורך הקלט.**

כאמור, הטרנספורמר בונה ייצוג של דאטה באמצעות ניתוח של **קשרים בין כל חלקיק הקלט המבוצע באמצעות מנגנון הנកרא (SA)** - הלב של הטרנספורמר. זאת אומרת, אנו צריכים לבצע חישוב עבור $O(M^2)$ זוגות של איברי הקלט עבור קלט באורך M. זה עלול להיות מאוד בעייתי מבחינה משאבי אחסון וזמן עיבוד הנדרשים לכך עבור תमונות ברזולוציה גבוהה (עקב מספר הפאטיים הגבוה). דרך אגב, בשנתיים האחרונות ייצאו מספר עבודות המציאות וריאנטים זולים יותר חישובית של הטרנספורמר כמו [Informer](#), [Reformer](#) ומאמר שסקרטטי לאחרונה [Performer](#) אולם למיטב ידיעתי, גרסאות אלה טרם הצליחו להשווות לרמת הביצועים של הטרנספורמר הקלסטי ב嚷גון מושימות.

תמצית מאמר:

כמו שהוסבר ב-[TransGAN](#) הסיבוכיות הריבועית של הטרנספורמר (למענה של מנגנון Self Attention) היא התוצאה של מכפלה (נסמן אותה ב-L) של מטריצות $X'Q=Q$ ומטריצת $X'K=K$ המשוחלתת כאשר 'Q', 'K', הם מטריצות Query ו-X Key ו-'Q' הוא מטריצה המייצגת קלט לטרנספורמר. הגודל של מטריצות Q ו-K הוא $M \times D$ כאשר M הוא אורך סדרת הקלט ו-D הוא מידת ייצוג הדאטה. מכאן קל לראות בבירור מאייה צזה הסיבוכיות של $O(M^2)$ של SA. נזכיר שהפלט של SA מחושב כ- LV , כאשר $X'V=V$ ו-'V' היא מטריצת Value.

להבדיל מרוב המאמרים המציעים גרסאות זולות חישובית של הטרנספורמר על ידי קירובים שונים לתוצאה של מנגנון SA, המאמר הנסקר מציע לתקוף את הבעיה מכיוון שונה לגמר. המאמר מציע ללמידה (!!) את מטריצת Q במקום לחשב אותה מהקלט. זה מאפשר לקבוע את הגודל של Q להיות הרבה יותר קטן מאשר אורך הקלט M, כך שסיבוכיות חישוב המכפלה של Q ב-K לא תהיה ריבועית-ב-M אלא (MN)O.

רעיון בסיסי:

המאמר מציע לחשב את Q بصورة $A'Q$, כאשר A היא מטריצה נלמדת, הנקראת מערך latent array. מטריצות V ו-K מחושבות بصورة מאוד דומה למנגנון SA המקורי. לאחר מכן במקום לחשב את הביטוי עבור Self-Attention הקלט X, המאמר מחשב את מה שנקרא Cross-Attention בין הקלט X לבין המערך הלטנטי A. גודל המערך הלטנטי A הרבה יותר קטן מגודל הקלט - וכך נמנעת הסיבוכיות הריבועית במונחי אורך הקלט.

הערה: מנגנון (CA) Cross-Attention הוזג לראשונה במאמר [BERT](#) ושימש לחישוב קשרים בין הפלט של האנקודר של BERT לבין פלטי ביןיהם של הדקودר במשימות כמו תרגום אוטומטי או Text Summarization.

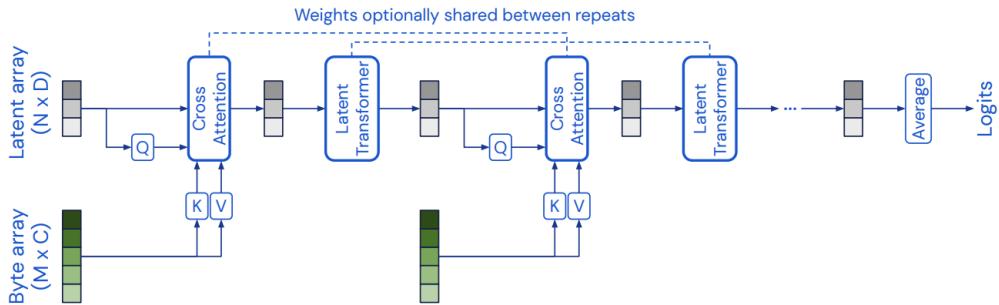


Figure 1. The Perceiver is an architecture based on attentional principles that scales to high-dimensional inputs such as images, videos, audio, point-clouds (and multimodal combinations) without making any domain-specific assumptions. The Perceiver uses a cross-attention module to project an input high-dimensional byte array to a fixed-dimensional latent bottleneck ($M \gg N$) before processing it using a stack of transformers in the low-d latent space. The Perceiver iteratively attends to the input byte array by alternating cross-attention and latent transformer blocks.

תקציר מאמר:

כעת נסביר את מבנה הקלטים למנגנון CA במאמר הנוסך. מטריות **K** ו-**V** נבנות בצורה זהה למנגנון SA המקורי - ככלומר באמצעות הכפלת הקלט במטריות **'V'** ו-**'K'** הנלמודות, בהתאם. מכיוון שאנו כבר לא מוגבלים עם הסיבוכיות הריבועית (במנוחה אורך הקלט) ניתן לחתך סדרת קלט ארוכה יותר מאשר בטרנספורמר הרגיל. למשל כאשר הקלט לטרנספורמר הוא תמונה ברזולוציה גבוהה, נהוג לחלק אותה לפאצ'ים בגודל 16×16 , בעוד שגבולת הסיבוכיות של הטרנספורמר המקורי. שימוש במערך לטנטי **A**, שנitin לבחור את גודלו לפי משאבי חישוב העומדים לרשותנו, מסיר מגבלה זו, המונעת מאייתנו להכניס לטרנספורמר סדרות קלט ארוכות. למעשה, המאמר מציע "לשתח" את הקלט ולהפוך אותו ל"מערך בתים" (byte-array) לפחות שמכפילים אותו במטריות Key ו-Value. אם הקלט הוא תמונה, כל איבר במערך הבטים מכיל את ערכו של הפיקסל (!!).

כמוון ניתן להכניס ל-Perceiver גם סדרות אודיו ארוכות או קטעי וידאו. יתרה מזו, המאמר טוען שניתן ל-Perceiver גם סדרות וידאו ייחד (!!)) עם אודיו במקשה אחת, דבר שלא היה אפשרי בגרסאות הקודמות של הטרנספורмерים (שדרשו התאמות לארכיטקטורה של הטרנספורמר בהתאם לסוג הקלט). ככלומר, הארכיטקטורה שהוצעה במאמר היא אגנוטיטית (!!)) לשוגים רבים של קלט זהה דבר חזק מאוד בעצמו.

ארQUITקטורה של Perceiver: פרטיים

לאחר שהבנו את העקרונות הבסיסיים של ארכיטקטורת Perceiver, ניתן לתאר את שאר הפרטים לגביה. לאחר חישוב של Cross-Attention בין המערך הלטנטי לבין הקלט, הפלט (של CA) מזון לטרנספורמר רגיל, הנקרא במאמר הטרנספורמר הלטנטי (LTr - latent transformer). חשוב לציין כי הגודל של הפלט של CA מזון CA אינו תלוי בגודל המקורי של הקלט אלא בגודל של המערך הלטנטי (הנקבע כאמור בהתאם למשאבי חישוב זמינים). מכיוון שהגודל של המערך הלטנטי בדרך כלל הרבה יותר קטן מגודל הקלט המקורי, ניתן "להעביר" אותו דרך LTr בסיבוכיות סבירה. ארכיטקטורה של LTr דומה לארכיטקטורה של 2-GPT ומורכבת מהדקודר של [המאמר המקורי](#).

הפלט של LTr שוב מזון CA בדומה למה שעשינו לפני כן (לשם כך משתמשיםשוב במטריות **K** ו-**V** המוחשבים מהקלט המקורי המשותח). הפלט של CA מזון ל-LTr-CA ועוד כאשר השילוב הזה (CA+LTr) יכול לחזור על עצמו פעמים רבות ליצור ארכיטקטורה עמוקה ועוצמתית המסוגלת לבנות ייצוגים חזקים לקלטים במספר דומיינים. נציין כי כל ה-LTr-CA יכולים להשתמש באותו משקלים (shared weights), משקלים שונים לכל אחד ZTr, או כל אופציית ביןיהם שהיא (למשל 3 סטיים של משקלים לכולם). ניתן לחשב על Perceiver כרשת מירונים רב שכבותית כאשר כל השכבה מורכבת מ-CA ו-LTr.

פינט האינטואיציה:

ניתן להסתכל על מערך הlatent כי סט של "שאלות נלמדות" לגבי הקלט. דוגמא של "שאלת" אפשרית יכולה להיות: תמדו את הקשרים בין פאטי' ק שבמרכז התמונה לכל הפאטי' בתוך פאטי' יותר, המכיל את ק. דוגמא של "שאלת" אפשרית יכולה להיות: תמדו את הקשרים בין פאטי' ק שבמרכז התמונה לכל הפאטי'ם בתוך פאטי' גדול יותר, המכיל את ק (בשכבה CA הראשונה). בשכבות עוקנות יותר של Perceiver המערך הלטנטי (השאלות) כבר תלוי בערכים המוחשיים בשכבות הנמוכות, ובודמה לרשותות קונבולוציה, מנוסות לשערק את הפיצרים היוצרים סמנטים של התמונה. ניתן גם לחשב על Perceiver-CNN רב שכבות (כאשר כל שכבה מקבלת את הקלט כולו).

קידוד מיקומי (positional encoding):

כמו שכבר ציינו בסקרים עיקריים הקודמות של מאמרינו בנושא הטרנספוררים, מנגנוני SA ו-CA הם אגנוטיים ולסדר איבריו בסדרות הקלט. כמו כן יציג איבר סדרת קלט, המופק באמצעות CA ו-SA, ישר ללא שינוי גם לאחר הפעלת פרמוטציה כלשהי על סדרה/ות הקלט. כמובן שמצב זה אינו סביר עבור תרחישים שיש בהם סדר אינהרנטי בין איברי סדרת הקלט (למשל שפה טבעית, תמונה, וידאו, אודיו ועוד).

כדי להבהיר למנגנונים של CA ו-SA את המידע לגבי מיקום של כל איבר בסדרה, מוסיפים לסדרת הקלט את מה שנקרא הקידוד המיקומי (PE). שטרכתו של PE היא לקודד מיקומו (היחס) של כל איבר בסדרת הקלט. עבור CA המאמר משתמש ב-PE דומה לזה שהוצע ב-BERT (המבוסס על פיצרי פוריה). לעומת זאת עבור מנגנון SA ב-LTr, המאמר משתמש ב-PE נלמדים.

הנושא של הקידוד המיקומי נדון בהרחבה במאמר (געשו בו כמה שינויים מעוניינים והמחברים ניסו לתת אינטואיציה לסייעת שיפור הביצועים).

הישגי מאמר:

המאמר השווה את הייצוגים המופקים באמצעות Perceiver עם מספר שיטות אימון self-supervised (מוסיפים שכבה לינארית לרשף המפיקה את הייצוג (המאומנת), מאמנים את המשקלים של שכבה זו ובודקים ביצועים) וגם עם שיטות SOTA supervised במספר דומיניים:

- תМОנות
- וידאו
- אודיו
- וידאו עם אודיו
- עניינקיודות

Perceiver: General Perception with Iterative Attention



Figure 2. We train the Perceiver architecture on images from ImageNet (Deng et al., 2009) (left), video and audio from AudioSet (Gemmeke et al., 2017) (considered both multi- and uni-modally) (center), and 3D point clouds from ModelNet40 (Wu et al., 2015) (right). Essentially no architectural changes are required to use the model on a diverse range of input data.

עבור כל הדומיינים Perceiver הצליח להציגם יותר טובים מכל שיטותervised שהם בדקו (כולל אלו שבסיסיים על הטרנספורמרים). נציין כי חלק מסוימות, ש-Perceiver "התגבר עליו", נבנו עבור דатаה מדומין ספציפי תוך ניצול התכונות האינהרנטיות של הדטה בדומיינים אלו (כמו ResNet בדומין של תמונות). עם זאת הביצועים של Perceiver בכל דומיין היו טיפה פחות טובים מהשיטות supervised המנצלים את התכונות של דטה בדומיינים אלו.

ResNet-50 (He et al., 2016)	76.9
ViT-B-16 (Dosovitskiy et al., 2021)	77.9
ResNet-50 (RGB+FF)	73.5
ViT-B-16 (RGB+FF)	76.7
Transformer (64x64)	57.0
Perceiver	76.4

Table 1. Top-1 validation accuracy (in %) on ImageNet. Methods shown in red exploit domain-specific grid structure, while methods in blue do not. The first block reports standard performance from pixels – these numbers are taken from the literature. The second block shows performance when the inputs are RGB values concatenated with Fourier features (FF) of the xy positions – the same that the Perceiver receives. This block uses our implementation of the baselines. The Perceiver is competitive with standard baselines on ImageNet while not relying on domain-specific architectural assumptions.

	Fixed	Random	Rec. Field
ResNet-50 (RGB+FF)	39.4	14.3	49
ViT-B-16 (RGB+FF)	61.7	16.1	256
Transformer (64x64)	57.0	57.0	4,096
Perceiver	76.4	76.4	50,176

Table 2. Top-1 validation accuracy (in %) on permuted ImageNet. “Fixed” = permuted with a constant permutation for all images over the dataset. “Random” = random, per-example permutation. Methods that make strong assumptions about the structure of 2D data fare poorly when this structure is removed. All methods receive identical input features (RGB+FF). We also show the receptive field of the input units for each model on the right, in pixels. Note that both Transformer and Perceiver have a global view of all inputs in each first layer unit. ResNet-50 starts with a 7x7 convolution, hence each unit sees 49 pixels, and ViT-B-16 inputs 16x16 patches, hence 256 pixels are seen by each first layer unit.

ג.ב.

מאמר מאד מעניין, מציע שיטה מגניבה להתגבר על הסיבוכיות הריבועית של הטרנספורמה. הארכיטקטורה המוצעת במאמר אגנוטיטית למבנה של קלט, ויכולת לשמש כמו שהיא לבניית “צוגי” דטה בדומיינים מגוונים.

Review 48: VAEBM: A symbiosis between autoencoders and energy-based models

פינת הסוקר:

המלצת קרייה ממייק: מומלץ לאוהבי מודלים גנרטיביים כמו VAE ו-Is-VAE להרחבת אופקים, אך לא חובה.

בahirot כתיבה: בינונית.

רמת היכרות עם כלים מתמטיים וטכניקות של DL/ML הנדרשים להבנת מאמר: נדרש רקע טוב בשיטות דגימה מתקדמות (דינמיקה של Langevin) והבנה טובה במודלים גנרטיביים.

ישומים פרקטיים אפשריים: יוצרה תמונות באיכות טובה יותר מ-StyleGAN אך עדין זה לא נראה באופן עקב מורכבותה.

פרטי מאמר:

lienek למאמר: [זמן להורדה](#).

lienek לקוד: לא נמצא בארכיב.

פורסם בתאריך: 09.02.21, בארכיב.

הוצג בכנס: ICLR2021.

תחומי מאמר:

- מודלים גנרטיביים.
- variational autoencoder (VAE)
- energy-based models (EBM)

כלים מתמטיים במאמר:

- Reparameterization trick
 - Langevin dynamics
 - markov chain monte-carlo - MCMC
 - התפלגות גיבס.
-

המציה מאמר:

המאמר מציע מודל גנרטיבי המשלב VAE עם EBM בשביל להנחת מהיתרונות של שניהם:

- היכולת של EBM ליצג התפלגות מורכבות בצורה מדויקת.
- היכולת של VAE לגנרט דגימות בצורה מהירה ויעילה.

השילוב של VAE ו-EBM נותן מענה לחולשות העיקריות של שתי השיטות האלו:

- EBM: דוגמה מאד איטית המגבילה שימוש בגישה זו רק לגינרוט תמונות בגודל קטן.
- VAE: יכולת מידול לא מדויקת של התפלגות הדטה המתבטא ביצירה של תמונות מטושטות.

המאמר מציע ארכיטקטורה, הנקראת VAEBM, המורכבת משני מרכיבים עיקריים: VAE ו-EBM. ארכיטקטורה זו מנצלת את היכולת של רכיב ה-VAE בשביל ללמידה את המבנה הכללי של המרחב הלטנטי מחד, כאשר רכיב ה-EBM בא "لتקן" את אי-הדיוקנים של רכיב ה-VAE ב"אזורים שאין בהם DATAה אמרטוי". במאמר טוענים ש-VAE מצליח לבנות קירוב יחסית טוב של התפלגות הדטה, לא נדרש מספר צעדים גבוה עבור עדכון הפרמטרים של EBM. בנוסף, שימוש ב-VAE מאפשר להציג את יכולת הדגימה של EBM ע"י רפרמטריזציה של המרחב הלטנטי. ולבסוף, לאחר ו-VAE כופה על המרחב הלטנטי להיות מפוגל עם התפלגות רציפה, הוא "משרה" התפלגות "חלקה" יותר גם של הדטה שהוא יוצר, שגורם לדגימה יותר יעילה עם MCMC.

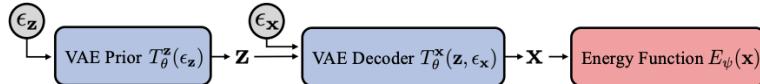


Figure 1: Our VAEBM is composed of a VAE generator (including the prior and decoder) and an energy function that operates on samples \mathbf{x} generated by the VAE. The VAE component is trained first, using the standard VAE objective; then, the energy function is trained while the generator is fixed. Using the VAE generator, we can express the data variable \mathbf{x} as a deterministic function of white noise samples ϵ_z and ϵ_x . This allows us to reparameterize sampling from our VAEBM by sampling in the joint space of ϵ_z and ϵ_x . We use this in the negative training phase (see Sec. 3.1).

ראיון בסיסי:

המאמר מגדר את ההתפלגות של המודל המגנרט (\mathbf{z}, \mathbf{x}) כמכפלה של $(\mathbf{z}, \mathbf{x})_{vae_k}$, ההתפלגות הרגילה של VAE ו- $(\mathbf{x})_{ebm_k}$ - פונקציית ההתפלגות הסטנדרטית של EBM (כלומר התפלגות גיבס). כאן \mathbf{x} היא דוגמה מהדומיין המקורי (למשל תמונות) ו- \mathbf{z} הוא וקטור לטנטי מרחב בעל מימד נמוך. כਮון שלכל אחד מהמודלים VAE ו-EBM הפרמטרים משליהם, והם המאוננים יחד במטרה למקסם את $-\log p(\mathbf{x} | \mathbf{z})$ על סט האימון. (\mathbf{z} היא פונקציית ההתפלגות של דוגמה \mathbf{x} מהמרחב המקורי המתקבלת לאחר מרגיניליזציה של המשתנה הלטנטי \mathbf{z} מ- $(\mathbf{x})_h$). המאמר מראה כי המיקסום הישיר של ביטוי זה מחיב דגימות מההתפלגות האפואוטריונית של $\mathbf{z} | \mathbf{x}$, כאשר \mathbf{x} מגנרט עם VAE, ויש להז סיבוכיות חישובית גבוהה (עקב שימוש ב-MCMC לדגימה מההתפלגות זו). נציין כי פונקציית המטרה של VAEBM היא סכום של פונקציית המטרה הסטנדרטית של VAE, המסווגת ב- vae_f , וזו של EBM, המסווגת ב- ebm_f . לאור זה המאמר מציע לבצע את המיקסום של שני שלבים:

- מיקסום של vae_f , כאשר הפרמטרים של ebm_f מוקפאים.
- מיקסום את רכיב הלוס של EBM כאשר הפרמטרים של VAE מוקפאים.

בנוסף בשביל להקל על הדגימה מ- (z, x) , הנחוצה בשביל שערוך הגרדיאנט של רכיב ה-EBM, המאמר מציע לעשות רפרמטריזציה **משותפת** של x ושל המשתנה הלטני z . בעצם גישה זו מונעת דגימה זו שלביות (דוגמים את z ואז את x) שעלולה להיות מאוד בעייתית מאחר ו-($z|x$) קשiosa להיות מרווחת באיזור מאד קטן -MC-MCMC בדרך כלל מתקשה להתמודד עם המצב זה.

תקציר מאמר:

קודם כל בואו נרענן את זכרונו וזכור מה זה בעצם VAE ו- EBM.

אוטו אנקודר וריאצינוי (VAE):

VAE הינו מודל גנרטיבי שהומצא ב-2014 ומהווה הכללה של אוטו-אנקודר רגיל (AE - AutoEncoder), המשמש להורדת מידע של הדטה. להבדיל מ-AE, ב-VAE ההתפלגות על המרחב הלטני מוגדרת מראש (למשל בתור התפלגות גאוסית סטנדרטית).

VAE מורכב משתי רשותות: רשות האנקודר ורשות הדקודר. המטרה של האנקודר היא לחשב את הפרמטרים של הייצוג הלטני עבור פיסת>Data x מהמרחב המקורי. המטרה של הדקודר הינה לשחזר את הדוגמא מייצוג הלטני שלה z .

از איך עובד EVAE? קודם כל מחשבים את הפרמטרים של הווקטור הלטני של דוגמא x באמצעות האנקודר, דוגמים מהתפלגות, המוגדרת על ידי פרמטרים הנ"ל, את הווקטור הלטני z . לאחר מכן מזינים את z לדקודר לשחזר של הדוגמא המקורי x . פונקציית הלוס של VAE מרכיבת מרכיבים:

- לוס השחזר: דיקוח השחזר של הקלט x , הנמדד באמצעות מרחק L_2 בין x לקלט של הדקודר.
- מרחק בין התפלגות המטרה על המרחב הלטני לבין התפלגות, המוחשבת באמצעות האנקודר (במרחק KL).

נציין כי לא ניתן לבצע גזירה ישירה של פונקציית הלוס של VAE לפי הפרמטרים של הווקטור הלטני, המחשבים על ידי האנקודר (גזירה זו נחוצה במהלך אופטימיזציה המשקלים (backprop) של רשות האנקודר). כדי להתגבר על קושי זה, משתמשים בטريق של רפרמטריזציה. במקרה לדוגמה המוגדרת על ידי הפלט של האנקודר enc_o , דוגמים מהתפלגות קבועה ולא תלואה בפרמטרים (בדרכן כל גאוסית סטנדרטית). לאחר מכן מפעילים טרנספורמציה (לינארית במקורה הגאוסי), המוגדרת על ידי enc_o על דגימה זו כדי לדמות דגימה מהתפלגות בעל פרמטרים enc_o .

מודלים מבוססי אנרגיה (EBM): גם EBM הוא מודל גנרטיבי, אך להבדיל מ-VAE, הוא משערר את פונקציית התפלגות של הדטה בצורה מפורשת (כלומר ממדל אותה על יד רשות ניורוניים). בשביל לגנרט דוגמאות חדשות באמצעות EBM, צריך לדגם מפונקציית התפלגות, המשוערת על ידי (נסמן אותו ב- ebm_k). בדרך כלל דגימה זו מתבצעת באמצעות אחד הוריאנטים של MCMC. עקב הסיבוכיות הגבוהה של שיטת דגימה זו, כרגע ניתן לגנרט עם EBM רק תמונות קטנות (עד 64×64).

אימון של EBM מתבצע באמצעות מקסום של פונקציית מטרה, שהיא התוחלת של הלוג של ebm_k על סט האימון (נראות מירבית). ebm_k מוגדרת ע"י התפלגות Gibbs שהיא אקספוננט שלילי של פונקציית האנרגיה (x), מוכפלת בקבוע נרמול (בשביל לאלץ את ebm_k להיות פונקציית התפלגות). EBM מאומן באמצעות GD, כאשר הגרדיאנט של פונקציית הלוס מורכב מהפרש של ערכי פונקציית האנרגיה E עבור דוגמאות ebm_k (השלב השלילי), ושל על ערכי E עבור דוגמאות מסט האימון (השלב החיווי). מכיוון שלא ניתן לדגם ebm_k ישירות, משתמשים באחד הסוגים של MCMC - בדרך כלל בדינמיקה של לנגווין (LD). LD הוא תהליכי איטרטיבי הבונה דוגמאות של ebm_k ע"י הuzzת דוגמאות בכיוון הגרדיאנט השלילי (לפי הדוגמאות x) של E במטרה לדוגם $-E$ במקומות שבהם $-E$ ערכים נמוכים - ככלומר באזורי בהם לפונקציית התפלגות ערכים גבוהים. נציין כי בשלב זה לוקחים את חוב הזמן באימון של EBM.

איך בעצם משלבים את EBM ו- VAE: המאמר מראה כי ניתן לחסום את פונקציית הלוס של VAEBM מלמטה על ידי הסכום של הלוס הסטנדרטי של VAE והlös של EBM, המסומן על ידי ebm_L . נציין כי האופטימיזציה של ebm_L כוללת שיעור של גרדיאנט של ebm_L לפי הפרמטרים של VAE (כי הדאטה \mathbf{x} מגונרט על ידי VAE). שיעור זה מאד כבד מבחינה חשובית כי הוא הכרוך בדוגימה מההתפלגות הפוסטרורית של דגימות מ-VAE. עקב כך המאמר מציע לבצע אופטימיזציה של ebm_L ו- vae_L לسورגים, דבר שמנוע את הצורך לגזר את ebm_L לפי הפרמטרים של רשת VAE. הרעיון השני של המאמר זה שימוש בטרייק של פרמטריזציה על \mathbf{x} - \mathbf{z} (המשתנה הלטנטית) **בו זמני**, מהלך זה מונע את הצורך לדוגם מההתפלגות המותנת $\mathbb{Z}|\mathbf{x}$ כי דוגמה כזו עלולה להיות בעייתית (הסביר בהרבה בפרק "רעיון בסיסי").

ካחד ההרחבות של תהליכי האימון המאמר מציע לבצע כמה איטרציות GD בשביב לקרב את פונקציית ההתפלגות של \mathbf{x} לפונקציית האנרגיה E אחריו שלב האופטימיזציה של VAE (מנסימם להביא למינימום את מרחק KL ביניהם). זה מזכיר את העדכון של הגנרטור ב-GAN $\text{GAN}_{\text{Wasserstein}}$.

הישגי המאמר:

המאמר משווה את ביצועיו של VAEBM מול מודלים גנרטיביים רבים מסוימים ומראה את עליונותו של VAEBM במונחי FID - inception score (IS) ובמונחי $\text{freshet inception distance}$ (FID) שמשנchez אותו ב-StyleGAN2 על CIFAR10 זה שיש לו ארכיטקטורה מורכבת בהרבה. נציין לחוב את הביצועים החזקים של VAEBM על StackedMNIST. כל תמונה ב-StackedMNIST היא שילוב של 3 תמונות של MNIST המקורי אż ייש 1000 מודדים. VAEBM מצליח לשחזר את כל המודדים להבדיל מכמה מודלים עדכניים של GAN.

המאמר גם משווה את יכולות דוגימה של VAEBM מול מודל גינרוט חזק denoising score matching ומציין כי VAEBM יעל יותר מפי 12 ממנו בהיבט זה כאשר איקות התמונות המגונרטות באמצעות שני מודלים אלו היא ד' קרובה (לפחות ויזואלית).

Table 1: IS and FID scores for unconditional generation on CIFAR-10.

	Model	IS↑	FID↓
Ours	VAEBM w/o persistent chain	8.21	12.26
	VAEBM w/ persistent chain	8.43	12.19
EBMs	IGEBM (Du & Mordatch, 2019)	6.02	40.58
	EBM with short-run MCMC (Nijkamp et al., 2019b)	6.21	-
	F-div EBM (Yu et al., 2020a)	8.61	30.86
	FlowCE (Gao et al., 2020)	-	37.3
	FlowEBM (Nijkamp et al., 2020)	-	78.12
	GEBM (Arbel et al., 2020)	-	23.02
Other Likelihood Models	Divergence Triangle (Han et al., 2020)	-	30.1
	Glow (Kingma & Dhariwal, 2018)	3.92	48.9
	PixelCNN (Oord et al., 2016b)	4.60	65.93
	NVAE (Vahdat & Kautz, 2020)	5.51	51.67
Score-based Models	VAE with EBM prior (Pang et al., 2020)	-	70.15
	NCSN (Song & Ermon, 2019)	8.87	25.32
	NCSN v2 (Song & Ermon, 2020)	-	31.75
	Multi-scale DSM (Li et al., 2019)	8.31	31.7
GAN-based Models	Denoising Diffusion (Ho et al., 2020)	9.46	3.17
	SNGAN (Miyato et al., 2018)	8.22	21.7
	SNGAN+DDLS (Che et al., 2020)	9.09	15.42
	SNGAN+DCD (Song et al., 2020)	9.11	16.24
	BigGAN (Brock et al., 2018)	9.22	14.73
Others	StyleGAN2 w/o ADA (Karras et al., 2020a)	8.99	9.9
	PixelIQN (Ostrovski et al., 2018)	5.29	49.46
	MoLM (Ravuri et al., 2018)	7.90	18.9

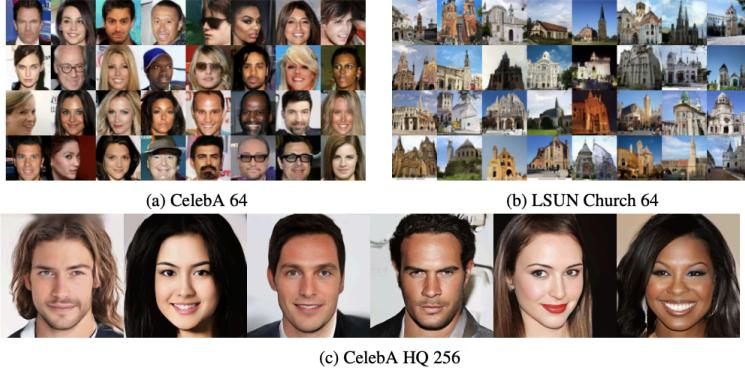


Figure 3: Qualitative results on CelebA 64, LSUN Church 64 and CelebA HQ 256. For CelebA HQ 256, we initialize the MCMC chains with low temperature NVAE samples ($t = 0.7$) for better visual quality. On this dataset samples are selected for diversity. See Appendix H for additional qualitative results and uncurated CelebA HQ 256 samples obtained from higher temperature initializations. Note that the FID in Table 3 is computed with full temperature samples.

דעתהוטים:

SVHN, CIFAR100, CelebA, StackedMNIST

ג.ב.

המאמר מציע הרעיון ד' חזק בתחום המודלים הגנרטיביים המשלב VAE ו-EBM ומראה תוצאות מבטיחות. בינהם עוד לא ברור האם רעיון זה יכול לסקן את שליטותם של GANs בתחום.

שנקרא:

Review 49: Exemplar VAE: Linking Generative Models, Nearest Neighbor Retrieval, and Data Augmentation

פינט הסוקרי:

המלצת קריאה ממייק: חובה רק למי שמתעניין Exemplar Models וגם מבין קצת ב- VAE - לאחרים ניתן להספיק בסקירה (:).

בahirot ctiyha: ביןנית.

רמת היכרות עם כלים מתמטיים וטכניקות של DL/ML הנדרשים להבנת מאמר: הבנה טובת בעקרונות VAE, ידע בסיסי ב-. kernel density models

ישומים פרקטיים אפשריים: יצירה של דוגמאות חדשות למטרת אוגמנטציה של נתונים קיימים למשימות שונות.

פרטי מאמר:

لينك למאמר: [זמן להורדה](#).

لينك לקוד: לא נמצא בארכיב.

פורסם בתאריך: 04.03.21, בארכיב.

הוצג בכנס: NeurIPS 2020.

תחום מאמר:

- [variational autoencoder - VAE](#)
- מודלים גנרטיביים לא פרמטריים שיוצרים דатаה "שירות מהדוגמאות של סט האימון" (generative models - EGM).

כלים מתמטיים, מושגים וסימונים:

- משערך חלון של פארזן ([parzen window - PW](#)) או שערוך צפיפות באמצעות קרNEL ([estimation - KDE](#)).
- חסם תחתון של [ELBO evidence](#).
- מרחק [KL](#) בין מידות הסתברות.
- מודלים של תערובת גאוסיאנים ([gaussian mixture models](#)).

מבוא:

המאמר מציע לשלב שתי גישות לצירת DATA (גינרטוט): VAE וגישה לא פרמטרית לצירת DATA ישירות מהדוגמאות מסט האימון, הנקראת EGM. שיטות משפחת EGM יוצרות דגימות חדשות ע"י בחירה באקראי של דוגמא מסט האימון והפעלת טרנספורמציה עליה. אחד היתרונות של שיטות אלו הינה הקלות של עדכון המודל: כאשר DATA חדש נוסף לדאטא סט, אין צורך באימון נוספת. החיסרון המשמעותי של גישה זו הוא הצורך בהגדרת מטrikaה במרחב DATA, הנדרשת להגדרת "סביבה של נקודת DATA". למידת מטrikaה כזו במרחבים בעלי ממד גבוה כמו בדומין היזואלי היא מאוד קשה. חיסרון נוסף של שיטות מסווג זה הוא הצורך לשמור את כל DATAהסט בשבייל ליצור דגימות חדשות שעלול להיות די יקר מבחינת גודל הזיכרון (עבור משימות מסוימות זה גם עלול להיות בעייתי בהיבט הפרטיטו).

לעומת זאת מודלים גנרטיביים פרמטריים לדוגמא VAE, GAN, זרימה מונרמלת (normalized flow) וגישות פרמטריות נוספות על רשותן ניירונים עמוקות ללמידה התפלגיות מורכבות למרחבים במימד גבוה. במודלים גנרטיביים פרמטריים רשת ניירונים מאומנת ליצור פיסות DATA חדשה ש"נראית טבעי" מדגימות של וקטורים אקראיים בעלי רכיבים בלתי תלויים (הנקראים הווקטורים הלטנטיים) מהתפלגות נתונה לא פרמטרית (!!). התפלגות זו (הנקראת התפלגות פרוירית) היא בדרך כלל (אך לא בהכרח) גaussית עם מטריצת קוריאנס איחידה וקטור תוחלות אפס. אחרי שהאימון הסטיים אין לנו צורך לשמר את סט האימון. אולם אם נרצה להוסיף DATA חדשות לדאטהסט, נצטרך סיבוב נוסף (צריך לציין שלבדיל מהטיבוב הראשון לא נעשה את האימון מופיע אלא נעשה סוג של ציול (fine-tuning) של המודל שהתקבל מהטיבוב הראשון.

המאמר מציע לשלב את שתי גישות אלה במטרה ליהנות מיתרונותיה של כל אחד מהם.

תמצית מאמר:

המאמר מציע לאמן VAE עם התפלגות הפרIOR (מעל המרחב הלטנטי) שהיא תערובת של גאוסיאנים (gaussian mixture) כאשר המרכז (וקטור תוחלות) של כל גאוסיאן הוא הייצוג הלטנטי של דוגמה מהDATAהסט. למעשה, ניתן לראות תערובת גאוסיאנים מעלה ווקטוריהם הלטנטיים של דוגמאות מהDATAהסט בתור משערץ צפיפות קרני (KDE) להתפלגות של המרחב הלטנטי של DATAהסט. ל-VAE בעל פרIOR זה (הנקרא Ex-VAE או Examplar VAE) להתפלגות של המרחב הלטנטי של DATAהסט. לא צריך לשמר את סט האימון. אולם אם נרצה להוסיף DATA בקצרה יש יתרון משמעותי מודלים גנרטיביים לא פרמטריים: לא צריך לשמר את הדוגמאות במרחב המקורי שלהם (מרחיב בעל ממד גבוה) וכן לסתופק רק בייצוגים הלטנטיים שלהם, שודושים הרבה פחות מקום אחסון. מצד שני כאשר עוד נספות נקודות לדאטא סט, לא מוכרים לאמן את המודל מחדש.

אצין שבמקרה זה הייתה עושים fine-tuning לשפת המקודדת (שbona kod letnati של דוגמא) מכיוון שהדוגמאות שנוספו עשויים לתורם בייצוגים הלטנטיים שהוא יוצרת. דרך אגב, ניתן לאמן את Ex-VAE על חלק מהDATA סט וליצור דוגמאות חדשות על שאר הדוגמאות (שלא השתתפו באימון).

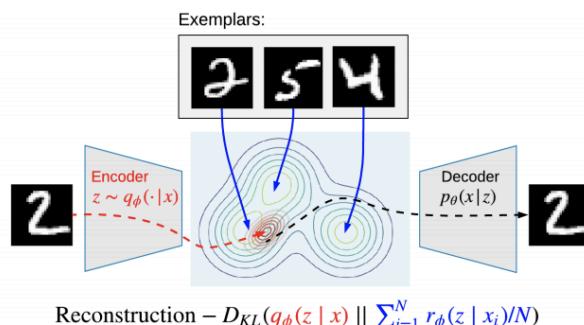


Figure 1: Exemplar VAE is a type of VAE with a non-parametric mixture prior in the latent space. Here, only 3 exemplars are shown, but the set of exemplars often includes thousands of data points from the training dataset. The objective function is similar to a standard VAE with the exception that the KL term measures the disparity between the variational posterior $q_\phi(\mathbf{z}|\mathbf{x})$ and a mixture of exemplar based priors $\sum_{n=1}^N r_\phi(\mathbf{z}|\mathbf{x}_n)/N$.

הסבר של רעיונות בסיסיים:

נתחיל את הדין מרגען של עקרונות VAE:

הסבר קצר על VAE:

VAE מורכב משתי רשתות נוירונים.

- הרשות המקודדת enc_N (דקודר) שבונה "יצוג" (קוטור) לטני של DATA. הקלט ל- enc_N הינו דוגמה x והפלט הינו פרמטרים (!!!) של התפלגות פוסטרIOR של הוקטור הלטנטי של x , כלומר הפרמטרים של $(x|z)P$. למשל אם התפלגות פוסטרIOR היא גaussית, הפלט של enc_N הוא קוטור התוחלות ומטריצת קוריאנס של קוטור ה"יצוג". לאחר מכן מגירים וקוטור z עם פרמטרים אלו (למעשה עושים זאת דרך טרייק של פרמטריזציה ולא ע"י דגימה ישירה).
- הרשות המפענחת dec_N (דקודר) מקבלת קלט וקוטור "יצוג" z והופכת אותו לדגימה מהמרחב המקורי. המטרה של הדקודר הינה לשחזר באופן כמה שיותר מדויק את הדוגמא x שמננו נוצר הקיים הלטנטי z .

פונקציית הלווי של VAE נגזרת מהחסם התחthing של evidence (נקרא ELBO) ומורכבת משני איברים:

1. לוס השחזר rec_L המשערך עד כמה טוב הצלחנו לשחזר את פיסת הדטה המקורי x .
2. מרחק KL בין התפלגות פרIOR $(z|z_k)P$ נתונה לבין התפלגות הפוסטוריית $(x|z)P$ הממודלת באמצעות הרשות המקודדת enc_N . המטרה של איבר זה הינה לכפות על $(x|z)P$ להיות קרובה ל- $(z|z_k)P$ - ניתן לראותו אותו כאיבר רגולרייזציה. כאמור $(z|z_k)P$ בדרך כלל נבחרת כ gaussית עם וקטור תוחלות אפס ומטריצת קוריאנס יחידה. הקירוב של $(x|z)P$ המחשב ע"י dec_N נקרא הקירוב הוריאצוני - נסמן אותו ב- $(x|z)q$.

מי שצריך הסבר יותר מפורט על VAE מוזמן להבitem ב- [פואט המעולה זהה על VAE](#)

הערות לגבי התפלגות הפרIOR והפוסטוריור של VAE:

ניתן לראות את $(z|z_k)P$ ב- VAE גם בתור "התפלגות יעד" בשביל $(x|z)P$. זה נובע מהעובדה שאחת המטרות של אימון VAE הינה מזעור של מרחק KL בין $(x|z)P$ ל- $(z|z_k)P$.

כאמור Ex-VAE מהווה הכללה של ה-VAE המקורי כאשר התפלגות הפרIOR $z|z_k)P$ הינה פרמטרית ומוגדרת כתערובת גאוסיאנים $(x|z)mix_P$ עם המרכזים ביצוגים הלטנטים של הדוגמאות. נציין כי לכל גאוסיאן בתערובת זו יש מקדם $N/1$ כאשר N זה מספר הדוגמאות (exemplars) המשמשות לאימון של Ex-VAE (המאמר מצין שלא חיבים להשתמש בכל הדאטasset לאימון).

פונקציית הלווי של Ex-VAE היא מאוד דומה לזה של VAE המקורי ומכליה שני איברים:

- לוס השחזר - זהה ל VAE
- מרחק KL בין הקירוב הוריאצוני של הפוסטוריור $(x|z)q$ לבין $(x|z)mix_P$. ברוח ההסביר הניתן בהערה לגבי הפרIOR והפוסטוריור, אחת המטרות של האימון היא "לכפות" על התפלגות הפוסטוריור להיות קרובה

כל האפשר לתערובת גאוסיאנים ($\mathbf{x}|z$)_P, המהווים שערוך קרנלי של הצפיפות של המרחב הלטני של הדאטסהט.

از איך מאמנים את Ex-VAE? קודם כל נציג Ci Ex-VAE מרכיב מ-3 רשותות נירונים:

- הרשות המקודדת הרגילה enc_N שהופכת דגימה מהדומין המקורי לוקטור הלטני שלה.
- הרשות המפענחת $zvar_N$ המיועדת לבניית קירוב וריאציוני של התפלגות הפוטטריוור ($\mathbf{x}|z$)_P. נציג Ci ($\mathbf{x}|z$)_P ממודלת ע"י גאוסיאן עם מטריצת קווריאנס אלכסונית כאשר כל איבר בלבד הינו פונקציה של \mathbf{x} (המודלת ע"י הרשות).
- הרשות המפענחת $dvar_N$, בעלת אותו המשקלים הנלמדים כמו הרשות $zvar_N$, המיועדת לשערוך של התפלגות תערובת הגאוסיאנים ($\mathbf{x}|z$)_P - "התפלגות יעד" עבור ($\mathbf{x}|z$)_P. למעשה $dvar_N$ משערכת את ($i|x|z$)_P עבור הדוגמאות i_x (קרובות ל- \mathbf{x} במרחב הלטני) המשמשות לבניה של פיסת DATA, נציג Ci ($i|x|z$)_P מודلت ע"י גאוסיאן עם אותו וקטור תוחלות כמו ($\mathbf{x}|z$)_P אך עם מטריצת קווריאנס קבועה אלכסונית.

תהליך האימון:

מכיוון ש-Ex-VAE הינו סוג של VAE קלאווי ופונקציית הלוס שלו דומה לזה המקורי של VAE אתמקד רק בהבדלים החשובים בין האימון של VAE ושל Ex-VAE.

1. נציג Ci החישוב של ($\mathbf{x}|z$)_{mix_P} עלול להיות כבד חישובי אם N (מספר הדוגמאות המשתתפים באימון של Ex-VAE) גבוהה. הסיבה לכך נעוצה בעובדה ש-($\mathbf{x}|z$)_{mix_P} הינו סכום של N גאוסיאנים ($i|x|z$)_P עבור דוגמא i_x ויש צורך לחשב ערך של כל אחד מהם. המאמר מציע לקחת רק את הדוגמאות היכי קרובות ל- \mathbf{z} במרחב הלטני מבחןת המרחק האוקלידי. מכיווןuai אפשר לדעת לאיזה דוגמאות הייצוג הלטני \mathbf{z} היכי קרובות בכל איטרציה של אימון, והמאמר מציע לשומר מערך של כמה דוגמאות קרובות מהאיטרציות הקודמות. מערך זה מתעדכן כאשר מתגלה דוגמא עם הווקטור לטני קרוב מפסיק ל- \mathbf{z} . שיטה זו נקראת במאמר (kNN השכנים היכי קרובים) אבל שימושה לבלא מתבצע קליסטוור אמיתי ככלשה במהלך האימון).

2. המאמר מציע לא להשתמש באיבר המתאים לדוגמא i_x מתערובת הגאוסיאנים ($\mathbf{x}|z$)_{mix_P}, כאשר מעדכנים את המשקלים של הרשותות לדוגמא i_x . לטענת המאמר זה מונע התכנסות לפתרונות טריוויאליים המרכזים מדי בוקטורים הלטניים של הדוגמאות מהדאטסהט.

Method	Dynamic MNIST	Fashion MNIST	Omniglot
VAE w/ Gaussian prior	-84.45 ± 0.12	-228.70 ± 0.15	-108.34 ± 0.06
VAE w/ VampPrior	-82.43 ± 0.06	-227.35 ± 0.05	-106.78 ± 0.21
Exemplar VAE	-82.09 ± 0.18	-226.75 ± 0.07	-105.22 ± 0.18
HVAE w/ Gaussian prior	-82.39 ± 0.11	227.37 ± 0.1	-104.92 ± 0.08
HVAE w/ VampPrior	-81.56 ± 0.09	-226.72 ± 0.08	-103.30 ± 0.43
Exemplar HVAE	-81.22 ± 0.05	-226.53 ± 0.09	-102.25 ± 0.43
ConvHVAE w/ Gaussian prior	-80.52 ± 0.28	-225.38 ± 0.08	-98.12 ± 0.17
ConvHVAE w/ Lars	-80.30	-225.92	-97.08
ConvHVAE w/ SNIS	-79.91 ± 0.05	-225.35 ± 0.07	N/A
ConvHVAE w/ VampPrior	-79.67 ± 0.09	-224.67 ± 0.03	-97.30 ± 0.07
Exemplar ConvHVAE	-79.58 ± 0.07	-224.63 ± 0.06	-96.38 ± 0.24
PixelSNAIL w/ Gaussian Prior	-78.20 ± 0.02	-223.68 ± 0.03	-89.59 ± 0.07
PixelSNAIL w/ VampPrior	-77.90 ± 0.02	-223.45 ± 0.02	-89.50 ± 0.13
Exemplar PixelSNAIL	-77.95 ± 0.01	-223.26 ± 0.01	-89.28 ± 0.12

Table 3: Density estimation on dynamic MNIST, Fashion MNIST, and Omniglot for different methods and architectures, all with 40-D latent spaces. Log likelihood lower bounds (nats), estimated with IWAE with 5000 samples, are averaged over 5 training runs. For LARS [2] and SNIS [36], the IWAE used 1000 samples; their architectures and training procedures are also somewhat different.

הישגיו מאמר:

המאמר משווה את יכולות דגימות הנוצרות באמצעות Ex-VAE בשלושה תרחישים הבאים:

1. שערוך ציפויות ההיסטבורות: המאמר מראה כי ההיסטבורות הממוצעת של הדגימות הנוצרות באמצעות Ex-VAE הינה גבוהה יותר מאשר של שיטות המתחروفות.
2. עברך נתונים מתייג, מאמנים את Ex-VAE ללא שימוש בתיאוגים. המאמר מראה כי עם Ex-VAE, הקלאסטרים של קטגוריות שונות למרחב הלטני, יותר מופרדים מאשר עם השיטות המתחروفות.
3. כאשר יוצרים דוגמאות חדשות עם Ex-VAE כדי להגדיל נתונים, המאמר מראה שיפור בBITSUMים במשימת סיווג בגין לגישות המתחروفות.

atasets:

.MNIST, Fashion-MNIST, Omniglot, CelebA

ג.ב.

מאמר נחמד עם רעיון למודל גנרטיבי שלא נתקלתי בו בעבר. מסקרן האם גישה כזו או השכלול שלה מסוגלת להתחזר באיכות התמונה עם SOTA בתחום זהה, כוללGANs. בנוסף אני מוכחה לראות מחקרים נוספים בנושא שיטות גנרטיביות לא פרמטריות.

Review 50: Language Through a Prism: A Spectral Approach for Multiscale Language Representation

פינת הסוקר:

המלצת קריאה ממ"ק: חובה לאנשי NLP.

bahiorot כתיבה: ביןוני פלאו.

רמת היכרות עם כלים מתמטיים וטכניקות של DL/ML הנדרשים לבנת מאמר: ידע בסיסי במודלים של NLP, הבנה בסיסית בשיטות ייצוג של וקטור בתחום התדר (התמרת פוריה או התמרת קוסינוס).

ישומים פרקטיים אפשריים: חקירה של תוכנות מבניות של מודלי NLP במגוון סקאלות.

פרטי מאמר:

[לינק למאמר: זמן להורדה.](#)

[לינק לקוד: זמן CAN](#)

פורסם בתאריך: 09.11.20, בארכ'יב.

הוזג בכנס: NeurIPS2020.

תחומי מאמר:

- חקר תכונות מודלי NLP عمוקות.

כליים מתמטיים, מושגים וסימונים:

- אנליזה ספקטרלית לגילוי של קשרים במגוון סקלנות בייצוג של טקסט (אמבידיגן).
- [התמרת קוסינוס דיסקרטית](#) (DCT), התמרת קוסינוס דיסקרטית ההופכית (IDCT).
- [מסנן מעביר גבויים](#) (HPF), [מסנן מעביר נמכרים](#) (LPF), [מסנן מעביר פס](#) (BPF).

תמצית מאמר:

שפות טבעיות מתאפיינות בתכונות מבניות כמו סקלנות החל מהרמה של מילה עד רמת הפיישה והמסמרק. בהקשר זה נשאלת השאלה האם המודלים, המבוססים על רשות הנירונים בתחום NLP, توفים את התכונות ההיררכיות אלו? האם ניתן "לשפר את ביצועי הרשות אם מאלצים אותה" לחזות את התכונות הללו? איך תכונות אלו משתנות בין מודלים, המאוננים למשימות שונות? המאמר הנסקר מנסה לתת מענה על השאלות האלו.

למעשה המאמר מציע שיטה לבחון תכונות וביצועי מודל NLP נתון בסקירה נתונה ע"י הורדוטן של כל הסקלנות האחריות מהמודל. למשל בשביל לבדק את ביצועי המודל בסקללת קצרה טווח (רמת מילה) לשימוש ספציפית, הם מאלצים את המודל "לא להשתמש" בסקללות ארוכות טווח (משפטים, פסקאות וכדומה). זה נעשה ע"י שימוש בטכניקות ספקטרליות מתחום עיבוד אותות המאפשרות לסנן (בתחום התדר) רק את התכונות בסקללה הנדרשת. כאן סקללות קצרות טווח (רמת מילה) מיוצגות ע"י תדרים גבוהים כאשר סקללות ארוכות טווח מיוצגות ע"י תדרים גבוהים יותר (נפרט על כך בהמשך).

השיטה המוצעת מסתמכת על הפעלה של מסננים ספקטרליים על אקטיבציות של נירונים בשכבות שונות של הרשת לאורך הטקסט (זה ימיד ה"זמן" שלנו !!). כמובן אם נרצה לבדוק עד כמה סקללה קצרה (מילה או שתיים, תדרים גבוהים) משפיעה על ביצועי מודל, מוסיפים למודל שכבה המפלטת החוצה את כל הסקללות הארוכות (תדרים יותר נמכרים). אם ביצועי מודל לא משתנים בצורה משמעותית כתוצאה מסוינן זה, המסקנה היא ש"תליות (סקאלות) ברמת מילה" חשובות יותר ותרלביצוע מוצלח של המשימה מאשר תלויות ארוכות טווח. כמובן במשמעות זו "מודול מספיק להתמקד בתליות קצרות טווח בטקסט" בשביל להציג ביצועים טובים.

טכנית זו מאפשרת לבדוק את התוכנות (מידע) הקשורות לסקאלה ולהפריד אותן מהתכונות הסמנטיות של קטורי ייצוג של טוקנים. בשביל להגעה להפרדה זו מוסיפים למודל שכבה המעביר חלקים שונים של קטורי ייצוג של הטוקנים (אמבידיניגס) דרך מסננים ספקטרליים שונים.

הערה: המאמר טוען שביעירון ניתן להוסיף שכבה נוספת (Shenkarat Prism) לא רק בתור השכבה האחורונה של הרשת, אך בפועל בכל הניסויים שהם עשו, הם הוסיפו את Prism אחרי שכבת האמבדיניגס של BERT. בעקבות זה אתייחס בהמשך רק לשינוי הספקטרלי של שכבת ייצוג הטוקנים (אמבידיניגס).

כמו שכבר אמרנו, המיקום של קטורי הייצוג בטקסט משחק תפקיד של מידע ה"זמן". בסוף מאמנים את הרשת עם שכבת Prism למשימות שונות. אז משווים את הביצועים של Prism עם רשת עם הרשת המקורית במשימה זו בשביל לבדוק האם הפרדה זו תורמת לביצועים.

הסבר של רעיונות בסיסיים:

בואו ננסה להבין איך בעצם עובדת שכבת Prism:

- חלוקה לסקאלאות (תדרים): מחלקים את הרכיבים של קטורי הייצוג לכמה תת-קבוצות. למשל אם יש לנו אמבדיניגס באורך 768 ואנו רוצים לבחון 3 סקאלות שונות, הרכיבים 1,..., 256 (קבוצת אינדקסים S_1) יהיו "אחראים" על הסקאלה ראשונה עם התדרים הגבוהים ביותר (ברמת מילה עד שתי מילים נגיד), הרכיבים 257,..., 512 (קבוצה S_2) יציגו את הסקאלה השנייה עם התדרים הבינוניים (ברמת " המשפט") , ו- 256 הרכיבים האחרונים S_3 י"שייכו" לסקאלה 3 של התדרים הנמוכים ביותר (ברמת "פסקה/המסמר").
- בניית של קטורי דגימות T לכל ניירון באמבדיניג: לכל אינדקס i בוקטור הייצוג על פני כל הטוקנים בטקסט, בונים וקטורי דגימות i_T. למשל עבור רכיב מסוים בוקטור הייצוג (נגיד במיקום 213) ובונים וקטורי דגימות 213_T המורכב מכל הרכיבים מס' 213 על פני כל הייצוגים של הטוקנים בטקסט.
- העברה של וקטורי i_T דרך DCT: מפעלים את התמורה קווינוס דיסקרטיות DCT (יפורט בהמשך) על כל וקטור i_T ובונים להם את הייצוגים הספקטרליים (בתchrom התדר). הייצוג הספקטרלי של וקטורי דגימות i_T יסמן ב i_F. נציין כי כל וקטורי דגימות עוביים יותר מהtransformación כלומר אם יש לנו 150 וקטורי i_T, אנו צריכים לבצע 150 DCTים (לכל אחד בנפרד). חשוב לציין שההميد של כל וקטור i_T שווה למספר הטוקנים בטקסט(!).
- שינוי ספקטרלי של וקטורי i_F: לכל וקטור i_F בוחרים את המסלון הספקטרלי שלו לפי האינדקס i. וקטורי i_F עם אינדקסים מקבוצה S_1 (ברמת מילה) יעברו דרך מסנן מעביר גובהים HPF, האינדקסים מקבוצה S3 יעברו דרך מסנן מעביר נמוכים ואינדקסים מקבוצה S2 יעברו דרך מסנן מעביר פס BPF (הסביר על איך עובדים המסלונים נמצוא בפרק הבא).
- העברה של וקטורי i_F המסוננים דרך התמורה קווינוס ההופכית IDCT: למעשה TIDCT מעבירה את הספקטרום המסונן של הייצוגים בחזרה לתchrom "זמן" (נזכיר שאצלנו מימד הזמן זה האינדקסים של האמבדיניגס לאורך הטקסט). נסמן את התוצאה של פעולה זו כ i_T. שעבור כל i הוקטור i_T בניי מכל הרכיבים במיקום i של וקטורי הייצוג המסוננים.
- אימון רגיל של רשת (BERT) עם שכבת Prism.

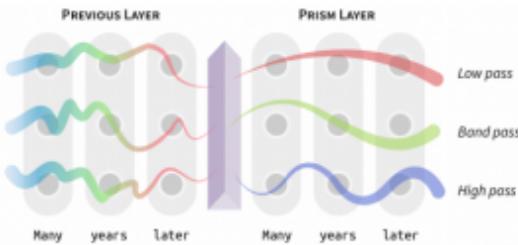


Figure 1: The prism layer specializes different neurons for different scales. First, the representations for an input are computed (left; in this case, the input is of length three). Next, a spectral filter (a low-, high-, or band-pass) is applied along the activations of each individual neuron (right). This produces neurons that are only able to represent structure at particular scales. Curved lines illustrate the scales at which neurons can change over an input.

הסבר בעניין התדרים:

השאלה המתבקשת כאן למה "סקליה של מילה" מייצגת דווקא תדרים גבוהים בזמן שה"סקליה של מסמן" מייצגת דווקא את התדרים הנמוכים ביותר? התשובה לכך נובעת מהעובדת ש"התדר של סקליה בטקסט" הינו ביחס הפוך ל"מחזור" של אותה סקליה. למעשה "המחזור" של "מילה" הינו נמוך ביותר בזמן של מחזור של "סקalias הפסיקה" הינו גבוה הרבה יותר. הסיבה לכך שהיא מורכבת ממספר מילימ, פחות משפטים ועוד פחות פסוקאות.

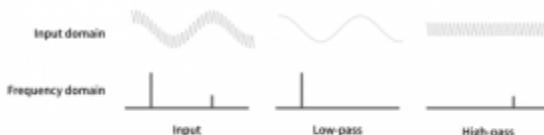


Figure 2: A visual depiction of spectral filters and their effects in the input and frequency domain. The input domain shows a sequence of values (e.g., the activation of a neuron across input tokens). The frequency domain shows the weight on the cosine waves which sum to produce the curve in the input domain. Low-pass filters only allow low frequencies to pass through, producing a smoothed input. High-pass filters only allow high frequencies and produce a locally-normalized input. Band-pass filters (not shown) are compositions of low- and high-pass filters.

הישגי מאמר:

בחינת "חשיבות" של סקaliasüber מודל עברו שימושות שונות: בשайл לבדוק את רמת ההשפעה של "סקalias" מסוימת" על ביצועי המודל, המחברים סיננו את כל הסקalias האחרות. נניח שאנו רוצים לבחון את ההשפעה של סקalias "המילימ" (תדרים גבוהים) על ביצועי מודל במשימה מסוימת. אז מפעילים מסנן שמסנן את כל התדרים האחרים (הנמוכים והבינוניים) על ידי העברת של ייצוג הטוקנים לאורק הטקסט דרך HPF בצורה המפורטת בסעיף הקודם. המאמר חילק את הסקalias (תדרים) ל-5 תחומי השווים באורך:

1. מילה - תדרים גבוהים.
2. פסוקיות (clause) - תדרים גבוהים-בינוניים.
3. משפט - תדרים בינוניים.
4. פיסקה - תדרים נמוכים-בינוניים.
5. מסמן - תדרים נמוכים.

Filter	Ex. Scale	Period (toks)	DCT index
HIGH	Word	1–2	130–511
MID-HIGH	Clause	2–8	34–129
MID	Sentence	8–32	9–33
MID-LOW	Paragraph	32–256	2–8
LOW	Document	256–∞	0–1

(a) **The spectral filters we consider in this work**, along with their periods, spectral bands (the indices in the DCT), and example linguistic phenomena at that scale. The period of a cosine wave for a DCT index is the approximate number of tokens it takes for the wave to complete a cycle.

ההבדיקות המוצגות במאמר עליה כי למשימת זיהוי נושא, התדרים הנמוכים הם הći חשובים שזה די הגיוני כי המודל צריך "להבין" את כך הטעטט ככל פחות או יותר בשביל להזוהה את הנושא שלו. מה שקצת מפתיע ב>Showcases שליהם זה השימוש המשמעותי בביטויים של המודל מול המקורו אחריו סינון של התדרים הגבוהים (סקאלה של מילה). במשימת סיווג אופי תגובה בדו-שיח, התדרים החשובים הם הבינוניים אבל לא בפער גדול על התדרים האחרים. במשימת זיהוי חלקן דיבור התדרים הגבוהים יוצרים הći שימושיים שזה די מובן בהתחשב לאופי המשימה. הררי בשוביל להבין לאיזה חלק דיבור לשיר מילה, מספיק לקחת בחשבון מילה או שתיים סמוכות.

מעניין שלמשימת זיהוי מילה ממוקמת שעלייה אומן BERT (בנוסף לזרחי סדר המשפטים) התדרים הכי חשובים הם הגבוהים ביותר בשוביל לנחש מילה "תחת מסכה" מספיק לדעת מילה או שתיים מסביב אליה. לעומת זו תגלית מאוד מסקרנת (!!).

ביצועי מודל עם שכבת Prism:

המחברים הוסיףו שכבת prism ל-BERT ובדקו את ביצועיו על 3 המשימות שתוארו בפיסקה הקודמת. הם הצליחו לשפר את הביצועים בצורה משמעותית לשתי משימות מתוך שלוש, כאשר עבר משימת זיהוי חלקן דיבור הם קיבלו תוצאות נמוכות טיפה מ-BERT המקורי. האימון בוצע על Dataset-WikiText-103.

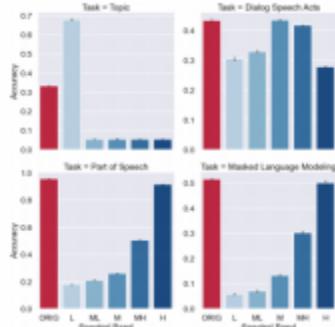


Figure 4: Different spectral filters extract information useful for tasks at different scales. Probing accuracy for different tasks and band-passes. A low-pass filter produces representations that yield highest probing accuracy on topic classification, while high-passed representations have highest probing accuracy for part of speech tagging. Meanwhile, band-passing the middle frequencies is most useful for dialog speech act probing. "ORIG" refers to the performance of the original token representations. Error bars show standard deviations over three probing runs.

הסבר על מושגים חשובים במאמר:

התמרת קוינינו דיסקרטית שלה DCT והופכית שלה IDCT: למעשה זה מקרה פרטי של התמרת פורייה הסטנדרטית. היא פועלת על סדרה של מספרים ממשיים ומבצעיה אותה לסדרה ממשית מאותו אורך בתחום התדר. אינטואיטיבית, התמורה זו מחפשת דמיון בין הסדרה לפונקציות קוינינו מתדרים שונים.

דאטאסתיטים ומשימות:

- משימת זיהוי אופי תגובה בדו-שיח: (Dialog speech act classification) השתמשו ב Switchboard Dialog speech acts corpus.
- משימת זיהוי נושא: 20 Newsgroups dataset.
- משימת זיהוי חלק דיבור: Penn Treebank.

ג.ב.

מאמר עם תוצאות מאוד מתקינות, המשמש בטכניות ספקטרליות לבחינה של תכניות (אורכי תלויות) עבור מודלי NLP עמוקים במשימות שונות. לצורך ביצוע הgisה המוצעת במאמר נבדקו על מנת משימות ורק על דאטאסט אחד בלבד לכל שימושה. עובדה זו קצת מקשה עליו להשתכנע שהתוצאות שהם גילו מתרחשים במשימות NLP אחרות בדאטאסטים אחרים. אני מצפה להמשך של המחבר המעניין הזה....

Review 51: Explaining in Style: Training a GAN to explain a classifier in StyleSpace

פינט הסוקר:

המלצת קריאה ממילק: כמעט חובה (לא חיבים אך ממש מומלץ).

bahiorot כתיבה: גבוהה.

רמת היכרות עם כלים מתמטיים וטכניקות של ML/DL הנדרשים להבנת מאמר: נדרשת היכרות די עמוקה עם עקרונות StyleGAN והבנה בסיסית במושגי Model Explainability.

ישומיים פרקטיים אפשריים: המאמר מאפשר לאחד פיצרים ויזואלים, הגורמים לשינוי המשמעותי ביותר בהתפלגות התוצאה של רשות הסיווג עבור תמונה זו.

פרטי מאמר:

lienek למאמר: [זמן להורד](#).

lienek לקובד: לא הצלחתי לאתר.

פורסם בתאריך: 21.04.27, בארכ'יב.

הציג בכנס: טרם ידוע.

תחומי מאמר:

- Model Explainability
- GANs

כליים מתמטיים, מושגים וסימונים:

- StyleGAN2
 - Path length regularization Loss
 - LPIPS
 - KL divergence
 - Style reconstruction loss
-

מבוא:

בשנים האחרונות רשתות ניירונים השתלו על עולם הראייה הממוחשבת. רובן המוחלט של תוצאות SOTA בmagic של שימושות ראייה ממוחשבת הושגו באמצעות שימוש ברשתות ניירונים. אולם פתרונות אלו, ובפרט רשתות ניירונים המשמשות לSiege שונות - עדין מהווים סוג של "קופסה שחורה", מבון שההחלפות של הרשת לא תמיד "ובוננות" לבני אדם. למשל, קשה לנו להסביר אילו תכונות ויזואליות של התמונה הובילו לשינוע כזה או אחר על ידי הרשת. כמובן אנו לא תמיד יודעים מה גורם לרשת המסייעת לזהות חתול בתמונה בהסתברות גבוהה: צורה של אוזניים, שפם חתולי או צורה של אף.

קיימות שיטות המנסות "להסביר" את הסינוג המופיע באמצעות הרשת (נקרא להן שיטות הסבר) עבור תמונה מסוימת על ידי "זיהוי" איזורים בתמונה, אשר המשפיעים באופן משמעותי על הפלט בשכבה الأخيرة של הרשת (שכבת הסינוג). ככלומר, שניים באיזורים אלו משנים באופן משמעותי את הסינוג, הנitin על ידי הרשת (קרי, ההסתברות הנחיצית עבור אחת מקטגוריות הסינוג). איזורים אלו נקראים מפות חום (heatmaps).

LAGISHOT CAN ALSO HAVE TWO LIMITATIONS:

- שיטות אלו מזיהות איזורים (אובייקטים) מקומיים של תמונה המשפיעים על החלטות הרשת באופן ניכר. אולם יכולתן של שיטות אלו לזהות "תכונות" (ATTRIBUTES - אטריבוטים) יותר גלובלית של תמונה כמו גדלים של אובייקטים שונים או צבעים, המשפיעות בצורה משמעותית על הסינוג, הינה מוגבלות.
- שיטות אלו מצליחות לזהות את האיזורים "החשובים לסינוג" של תמונה אך לא מספקים אינדיקציה איזה שינוי באיזורים הללו נדרש לצורך להביא לשינוי כזה או אחר של פלט הרשת.

שיטת הסבר משפחת counterfactual explanations CE מtaggerates על קשיים אלו באמצעות זיהוי תכונות (להבדיל מאייזרים, אטריבוטים) של תמונה, המשפיעות באופן ניכר על פלט הרשת. זיהוי זה נעשה באמצעות ניתוח של פלט הרשת עבור תמונה, השונות מתמונה נתונה בכמה תכונות בודדות בלבד. לבסוף נבחר מספר קטן של תכונות המשפיעות באופן מוסף על הסינוג הנitin ע"י הרשת.

באופן טבעי שיטות CE בתחום היזואלי עושות שימוש נרחב ב-GAN-ים, הידועים ביכולתם ליצור תמונות מוקטור "פיצ'רים חכמים" z בעל מידת נמוך הרבה יותר מהתמונה. כדי לזהות אטריבוטים של תמונה אשר "חשובים" לדוחה של קטגוריה מסוימת, ניתן "לשחק" עם רכיביו של וקטור z כדי לראות אילו מהם משפיעים על פלט הרשות עבור קטgorיה זו באופן המשמעותי ביותר. חשוב לציין כי כאשר "הפייצ'רים" (הרכיבים) של וקטור z הם מעורבים (כלומר כל רכיב של וקטור "אחראי" על קומבינציה מסוימת של אטריבוטים ויזואליים של תמונה), קשה "לבודד" אטריבוט ויזואלי המשפיע ביותר על פלט הרשות.

תמצית מאמר:

כידוע StyleGAN היה אחד הגאנים הראשונים שהצליח "להפריד" (disentangle) את הפיצ'רים של וקטור קלט z כך שכל תטא-קבוצה של רכיביו הינה "אחראית" על פיצ'ר ויזואלי מסוים של תמונה (כגון צבע שיער ועיניים, אורך שיער, גוון של עור). יותר ספציפית, בשלב הראשון וקטור קלט של StyleGAN מושן לרשות, המפיקת ממנו את הפיצ'רים היזואליים המופרדים (הפלט של רשות זו נקרא וקטור סגנון - style vector). עקב כך StyleGAN הופך לכלי עזר טבעי לבנייה של "מסביר החלטות של הרשות", המבוסס עם אטריבוטים ויזואליים.

המאמר הנזכר מציע שיטה, הנקראת ExStyleGAN2, שבליה נמצאת StyleEx, שבליבה נמצא פיצ'רים ויזואליים של תמונה, המשפיעים ביותר על החלטה המופקת על ידי רשות מסווגת נתונה. נציין כי אכן, המאומן על דאטהסט נתון של תמונות, אינו בהכרח "יתפוא" פיצ'רים ויזואליים רלוונטיים למסורת נתון. למשל אכן, המאומן על דאטהסט תמונות של מכוניות עשוי שלא לגלוות פיצ'רים משמעותיים למסורת של דגם של מכוניות. כדי להתגבר על קושי זה, המאמר משלב את המஸוג המאומן באימון של StyleGAN2. כך StyleGAN2 המאומן לומד "להפיק" את הפיצ'רים היזואליים הרלוונטיים למסורת נתון. לאחר מכן, עבור כל קטגורית סיווג ובורחים את כל התמונות u_k המסוגות c -ה. בשלב האחרון מעתירם קבוצה של כמה פיצ'רים (style coordinates) שהשינוי בהם מקטין את ההסתברות המומוצעת של u מעל u_k .

הסבר של רעיונות בסיסיים:

כמו שתואר בפרק הקודם, האימון של ExStyleGAN מורכב משני שלבים. כתע נתאר כל שלב בצורה מפורשת יותר:

שלב 1 : אימון משותף של StyleGAN2 ביחד עם רשות מסווגת נתונה C

נזכיר כי כדי לאמן גאן, אנו מאמנים יחד את שתי רשותות:

- רשות הגרטטור G , שמטרתה ליצור תמונה מוקטור (זה יכול להיות וקטור גאומטרי או וקטור קבוע במקרה של StyleGAN שבו האלמנטים האקראיים "מזרקרים" ישרות לשכבות של G).
- רשות הדיסקרימינטור D , המאומנת כדי להבחן בין דוגמא מסט האימון לדוגמא מלאכותית שוגרתה על ידי G .

נזכיר שגם מעוניינים "לעצב" את מרחב הפיצ'רים (הנקרא מרחב הסגנון ל-StyleGAN) כך שיכלול אטריבוטים רלוונטיים לרשות מסווגת נתונה. דרך אגב, ה"עיצוב" של מרחב הסגנון המקורי מתבצע באמצעות של טרנספורמציה אפינית (הנלמדת) של מרחב הסגנון המקורי של StyleGAN2. המאמר מציע את השינויים הבאים לStyleGAN2- C :

- הוספה רשת מקודדת (Encoder) המיעדת לבנייה של וקטור סגןון מתמונה. הרשת המקודדת E מאומנת יחד עם G תוך כדי מזעור של לואש השחזר (תמונה מוגנת ל-E ולאחר מכן G משוחרת אותה והלואש מודד עד כמה טוב האלגוריתם לשחזר את התמונה). לעומת זאת E-G ייחד מהוות אותה.
- הגנרטור G מקבל את קלט גם את הסיווג עבור התמונה שהוא מייצר. תוספת זו מאפשרת להכין פיצרים, רלוונטיים לסיווג, למרחב הסגןון החדש.
- "לואש הסיווג" הוסיף לפונקציית לואש של StyleGAN2. לואש זה מודד מרחק בין הסיווג של התמונה המקורי (הקלט ל-E) לבין הסיווג עבור תמונה, המוגנרטת באמצעות G מהקלט של E. מרחק זה בין הסיווגים נמדד על ידי KL-divergence.

מבנה של פונקציית לואש עבור שלב 1:

פונקציית לואש מורכבת מ-4 איברים:

1. הלואש האדברסרי (הלוגיסטי) הרגיל של אן [מהמאמר המקורי של al et al](#).
2. הלואש שמטרתו לגרום לכך שכל שניי של וקטור הסגןון יביא לשינוי פרופורציוני בהתמונה הנוצרת. לעומת זאת, קטע בוקטור הסגןון צריך לגרום לשינוי קטן בתמונה הנוצרת ממנו וככל שוקטור הסגןון משתנה יותר, השינוי בתמונה הנוצרת ממנו יהיה גדול יותר. כהעתה אגב, לואש זה הוצע לראשונה ב-[StyleGAN2](#).
3. לואש השחזר שהסביר לעיל מרכיב מ-3 המוחברים הבאים:
 - a. איבר המודד מרחק L1 בין התמונה המקורי לתמונה המשוחזרת (המתקבלת באמצעות העברתה של התמונה המקורי דרך המקודד E ולאחר מכן דרך הגנרטור G).
 - b. איבר [LPIPS](#) המודד מרחק perceptual בין התמונה המקורי לתמונה המשוחזרת. איבר זה מודד מרחק בין "יצוגי" התמונות הללו המתקבלים באמצעות רשותות מסוימות כמו VGG או SqueezeNet.
 - c. איבר המודד מרחק L1 בין וקטורי הסגןון של התמונה המקורי לתמונה המשוחזרת. איבר זה הוא גרסה של לואש הסגןון שהוצע ב-[StarGANv2](#).
4. מרחק KL בין הפלטים של הרשת המסוגת עבור התמונה המקורי והתמונה המשוחזרת.

שלב 2: איתור של אטריבוטים ויזואליים "המשמעות" על הסיווג לקטגוריות

המטרה של שלב זה היא לאתר את הכוונים במרחב הסגןון הגורמים לשינויים משמעותיים בפלט של רשת הסיווג. חיפוש זה מתבצע באופן הבא:

- לכל קטגורית סיווג בוחרים את כל התמונות המסוגות עם קטgorיה זו על ידי הרשת המסוגת.
- מבצעים חיפוש של מספר (שנקבע מראש) קואורדינטות של וקטור הסגןון הגורמים לירידת הבולטת ביותר של ההסתברות המוגעת של קטgorיה זו (המחושבת על ידי הרשת המסוגת).
- לכל קואורדינטה של וקטור הסגןון מוצאים את הכוון של וקטור הסגןון (+1 או -1) הגורם לירידת הסתברות של קטgorיה זו.

נצין שניתן לישם שיטה זו גם למציאת של הכוונים "המשמעות" ביותר של תמונה נתונה.

הישג המאמר:

חייב להודות שהתוצאות לא פחות מפרשיות, לפחות מבחינה ויזואלית. ExStyleEx הצליח לזהות פיצרים ויזואליים לא מעורבים קלים להבנה עבור מגוון רשותות מסווגות בתחומיים מגוונים. למשל, הגישה המוצעת הצליחה לזהות פיצרים המשפיעים ביותר על זיהוי גיל, מין וגם על זיהוי מחלות של עיניים ואףלו מחלות של עליים.

ג.ב.

מאמר מגניב עם תוצאות מרשים שבבסיסו רעיון הגיוני וקל להבנה בתחום ה-explainability של רשותות ניירונים.

Review 52: Neuron Shapley: Discovering the Responsible Neurons

פינט הסוקרי:

המלצת קריאה ממילק: כמעט חובה (לא חייבים אף ממש מומלץ).

בahirot כתיבה: בינוי פלאס.

רמת היכרות עם כלים מתמטיים וטכניקות של DL/ML הנדרשים להבנת מאמר: היכרות בסיסית עם שיטות explainability כמו SHAP והבנה של מושגים סטטיסטיים בסיסיים כמו רוח סמך.

ישומים פרקטיים אפשריים: זיהוי ניירונים המשפיעים ביותר על ביצועי רשות.

פרטי מאמר:

لينק למאמר: [זמן להודהה](#).

لينك לקוד: [זמן כאן](#) (לא رسمي)

פורסם בתאריך: 20.11.13, בארכיב.

הוזג בכנס: NeurIPS 2020.

תחומי מאמר:

- חקר התנהגות של רשותות, ניירונים מאומנות, תורת המשחקים.

כלים מתמטיים, מושגים וסימונים:

- ערכי SHAP.
- שיטת מונטה קרלו לדגימה.
- בעיות שודדי מרובי ידים.
- רוח סמך (confidence interval).
- חשיבות של פיצ'רים (feature importance).

תמצית מאמר:

המאמר מציע שיטת למדידת תרומה של ניירון נתון על רשות נוירונים מאומנת על ביצועי הרשות. הרעיון בגודל הוא פשוט מאוד: אם איפוס של ניירון גורם לירידה משמעותית בביטויים של רשות הנוירונים, החשיבות (תרומה) של ניירון זה היא גבוהה, אחרת היא נמוכה. במידה מסוימת זה מזכיר "חשיבות של פיצ'" (feature importance) רק שכן אנו בוחנים את הפיצ'רים של המודל עצמו ולא את התוכנות של הקטל. המחברים הגיעו דומה לערכי SHAP הקלאסיים, שהפכו לאחרונה לאחד הכלים הפופולריים בשערוך חשיבות הפיצ'רים, ככלי למדידת חשיבות של נוירונים. באופן לא מפתיע "חשיבות של ניירון" נקראת במאמר ערך שאפל' של ניירון (Neuron Shapley - נקרא לזה N-Shapley בהמשך).

از מה זה בעצם ערך N-Shap? למעשה ערך N-Shap של ניירון i_N מודד את התרומה המומוצעת לביצועי הרשות, מושגת ע"י הוספת ניירון i_N לכל תת-הרשומות של רשות N , שלא מכילות את i_N . לעומת זאת כל תת-רשת של הרשות המאומנת N , מודדים את הביצועים שלה ואז מוסיפים לכל אחת מהם את i_N , שוב מודדים את הביצועים ובוסף מחשבים את ההפרש בין הביצועים של רשותות אלו. נציג שאנו לא מאמנים את תת-הרשומות אלא רק מודדים את הביצועים שלהם על דאטasset נתון. לעומת זאת ערך i_N -Shap של ניירון i_N מוגדר כממוצע של הפרשי הביצועים עבור כל תת-רשתות של N . שימו לב שבנוסחה (1) במאמר, המגדירה את i_N -Shap באופן פורמלי, מופיעות מקדמיםBINOMIAL המשמשים לחישוב של מספר תת-רשתות בגודל S .

כידוע המספר הכלול של תת-הרשומות של רשות נוירונים הינו אקספוננציאלי במונחי מספר הנוירונים ברשות. לכן גישה זו אינה ישימה אפילו עבור רשותות לא גדולות במיוחד (מאות אלפי נוירונים). כדי להתגבר על בעיה זו מחברי המאמר מציעו שתי גישות:

- **גישה מונטה-קרלו:** עבור כל ניירון i_N , דוגמים מספר תת-רשתות M (למעשה מגירים את הנוירונים המרכיבים רשותות אלו) באופן רנדומלי, לעומת זאת מקבלת הסתברות שווה להיבחר. אז SHAP- i_N של כל ניירון זה בעצם ממוצע של כל התורומות של על כל תת-רשתות שנדרמו עבורה. מכיוון שמספר תת-רשתות הינו אקספוננציאלי במספר המשקלים ברשות הגישה זו לא יעילה עקב השונות הגדולה של האומדנים של i_N -Shap המוחשיים באמצעותה (כאשר מספר הדגימות M הינו הרבה יותר קטן מאשר המספר הנוירוני הכלול num_N).

- **גישה אדפטטיבית המבוססת על הכלים מעולם MAB:** המאמר מצין כי למעשה אנו מעוניינים לאתר K נוירונים בעלי ערך i_N הגבוהים ביותר. עם ניסוח זהה הבעייה הופכת דומה לבעיה הקלאסית בתחום של MAB קרי מציאות "מוכנות הימורים בעלת הסתברות זכייה מקסימלית". ניתן לראות כי בעיה זוiskaola למציאה של K משתנים מקרים בעלי תוחלת הגבוהה ביותר מתוך סט גדול של משתנים אקראיים. בעיה זו נדונה באופן נרחב בספרות של MAB.

בתבסס על הבדיקה זו המאמר מציע אלגוריתם הנקרא (truncated MAB, Shapley T-MAB-S) שעבור K נתן מזהה K נוירונים עם התרומה הגבוהה ביותר. בגודל בכל איטרציה, עבור כל ניירון דוגמים תת-רשת אחת מחשבים את תרומתו עבור תת-רשת זו וمعدכנים את הממוצע, השונות ורוחה הסמך של ניירון זה. לאחר מכן מצמצמים את סט הנוירונים הנדגמים ע"י הוצאת נוירונים שרוחם סמך שלהם של תרומתם לא מכיל את ערך

התרומה ה-K המקסימלי (k-th largest) עברו אותה איטרציה. תרומת הנירונים שהוצאו (התוחלת והשונות) נשארת קבוע למשך כל האיטרציות הבאות. האלגוריתם עוצר כאשר לא נותרו נירונים בסט הנדגם (התהילר והאינטואיציה יפורטו בפרק הבא).

הסבר של רעיונות בסיסיים:

פריטים או אינטואיצה של האלגוריתם S-MAB-T:

- מגדירים את סט הנירונים הנדגמים U כסט המכיל את כל הנירונים של רשת N.
- עבור כל נירון π_N האלגוריתם דוגם תת-רשת אחת ומודדים את התרומה של π_N עבור תת-רשת זו. נציג שם הביצועים של לחת-הרשת שהוגירה, הם מתחת לספ (הנקבע מראש), תרומתו באיטרציה זו נקבעת לפחות.
- אחרי כל איטרציה מחשבים את ממוצע, שונות ורווח-סמרק של ערכי Shap_N עבור כל הנירונים מ-U, בהסתמך על הערכים שהתקבלו באיטרציות הקודומות. נזכיר כי רוח סמרק נבנה סביב הממוצע ורוחבו נמדד במספר שוניות סביב התוחלת (ראה [הסבר על בניית רוח סמרק](#) ליותר פרטים).
- מחשבים את הערך k-th המקסימלי $\text{Max}_K \text{Shap}_N$ שהתקבלו באיטרציה זו.
- מושאים מ-U את כל הנירונים Max_K לא שיר לחוש סמרק שלהם (עם איזשהו מרג'ין קטן משני הצדדים). ערכי Shap_N של נירונים אלו נותרים ללא שינוי למשך איטרציות הבאות.
- עוזרים כאשר סט הנירונים הנדגמים נהיה ריק.
- ובחרים K הנירונים עם ערכי Shap_N המקסימליים.

פינת האינטואיציה: למעשה Max_K הינו אומדן של מקסימום ה-K של כל ערכי Shap_N שנדגמו. כאשר אנו מושאים את הנירונים, שעבורם Max_K לא שיר לחוש סמרק שלהם (האינטרול שבו ערך Shap_N של נירון טופ-K אמור להימצא בהסתברות גבוהה), אנו מושאים את הנירונים [שהסבירות שערך \$\text{Shap}_N\$ שלהם יהיה בין טופ-K הינה נמוכה](#). כך מצמצמים את מספר הנירונים הנדגמים ע"י הוצאותם של "מועדדים לא טובים להיות בין טופ-K".

תכונות של Shap-N: כתעណון בשלוש תכונות הבסיסיות של מטריקת Shap-N:

- ערך Shap_N אףו לנירון π_N שקול לכך שהוסתו לכל תת-רשת לא משפייע בכלל על ביצועי הרשת.
- אם התרומות של שני נירונים לכל תת-רשת אפשרית (שלא מכילה את שני נירונים אלו) הין שוות, אז ערכי Shap_N של נירונים אלו שוויים גם כן.
- אדיטיביות:** נניח שיש לנו שני נתונים שכחישבנו עליהם ערכי Shap_N של נירון כלשהו. ניתן לראות כי ערך Shap_N עבור נירון זה המחשב על איחוד נתונים שונים אליו יהיה שווה לסכום של ערכי Shap_N שלהם.

בזכות תכונות אלו (שהמאמר הנסקר מוכח בצורה ריגורוזית), נטען במאמר כי N-Shap מהויה מטሪקה "טובה והגיונית" למדידה של תרומת נירון לביצועי רשות (אני חשב ש-N-Shap הינה מטሪקה טוביה בהקשר המדבר כי היא מהויה הרחבה טבעיות של ערכי שאפלி קלאסיים לרשותות נירוניים).

הסבר על מושגים חשובים במאמר:

ערך שאפלி: ערך שאפלி הינו כל קלאסוי לשערור של חשיבות של פיצרים בהינתן מודל מסוים. למעשה עושים משהו מאוד דומה לנעשה במאמר הנסקר - מודדים את השינוי בביצועים המתkeletal ע"י הוספה של פיצ'ר f לכל תתקבוצה של פיצרים (כאשר יש מספר רב של פיצרים משתמשים בקירובים בצורה דומה לממה שנעשה במאמר).

תיאור קצר של בעיתת "שודד מרובה ידיים" (MAB): נניח שיש לנו N מכונות מזל שלכל אחת יש הסתברות שונה לזכיה והסתברויות אלו לא ידועה למהמר. המטרה העיקרית בעיות MAB הינה (בഗדייל מאד) למקסם את הרוחות הממציע של המהמר ([הסבר על בעיתת MAB](#)).

הישגיו מאמר:

המאמר מראה כמה תוצאות מעניינות ודי לא צפויות לגבי ההשפעה של נירונים טופ-K על ביצועים המודל (עבור רשת V3/Inception7ו שאומנה על Imagenet). למשל המאמר מראה כי הוצאים של 10 נירונים בלבד (למענה זה איפוס של 10 קרנלים שמחוברים אותם) גורמת לירידה של 50% (!!!) בדיק כאשר האיפוס של 20 נירונים-caלו מרסק את הביצועים ל-8% (!!)). עוד דבר מעניין שהמחברים מצאו: אם מוציאים את הנירונים החשובים ליזויו של קטגוריה ספציפית, הדיק של קטגוריה זו מתרסך ואילו הפגיעה בבדיקה בקטגוריות האחרות היא די קתנה. צריך לציין שהמסקנות הללו הן לא אינטואיטיביות כלל (לפחות מבחינתי) – הררי כארן מאמנים רשות עם דרופאוט חשיבות של כל נירון בודד נתה להיות לא גבוהה במיוחד. לא הייתה משער שההורדה של 20 נירונים בלבד תוביל ל垦יסה מוחלטת של ביצועים.

בנוסף המאמר בדק מהם הנירונים "הכי רגילים להתקפות אדוורסריות", כלומר האם ניתן להציגן נגד התקפה נתונה באמצעות "אייפוס" של נירונים מסוימים. נזכיר כי התקפה אדוורסרית מנסה להנדס שינויים קלים ולא גראים לעין לתמונה במטרה לגזור לרשף לשנות את החיזוי של התמונה באופן שימושתי. המחברים מצאו כי אייפוס של נירונים עם התרומה הכי גבוהה בהקשר זהה מצליח לנטרל את התקפה כמעט לגמרי ואילו הביצועים של הרשות על הדוגמאות הרגילות ספגות ירידת קלה בלבד. שימו לב שגישה זו אינה מהווה דרך טובה להציגן נגד התקפות אדוורסיות. אייפוס נירונים הכי חשובים (בהקשר זה) מעניק הגנה נגד התקפה הספציפית בלבד (!!!) ואני דין ד' בקהלות לבנות התקפות דומות אחרות נגד רשות עם "הנירונים המאופסים". נראה שהתקפה החדשה תבחר נירונים אחרים בשביל "להתמקד עליהם". מעניין שהנירונים בעלי התרומה הכי גבוהה בהקשר האדוורסרי והנירונים בעלי ערכי N-Shap הגבוהים ביותר מושם הסיווג המקורי, יצאו ד' שונים.

ג.ב.

מאמר מעניין המשלב שיטות מתחום MAB וערך שאפלி לאנליזה של "מה שקרה בתוך רשותות נירונים מאומנוות". התוצאות של המאמר לא כל כך אינטואיטיביות והייתי שמח לראות עוד מאמרים בודקים את הסוגייה זו על יותר מושגים וארQUITקטורות רשות אחרות.

Review 53: Learning to summarize from human feedback

פינת הסוקר:

המלצת קריאה ממילא: מאוד מומלץ.

בahirot כתיבה: גבואה מינוס

רמת היכרות עם כלים מתמטיים וטכניקות של DL/ML הנדרשים להבנת מאמר: הבנה טובה בשיטות הקיימות של reinforcement learning, abstractive summarization וידע בסיסי ב-*reinforcement learning*.

ישומים פרקטיים אפשריים: אימון של מודלים לתמצאות אבסטרקטיבי עם עם פחות דата מתויג

פרטי מאמר:

lienק למאמר: [זמן להורדה](#).

lienק לקוד: [זמן כאן](#)

פורסם בתאריך: 27.10.20, בארכ'יב.

הציג בכנס: NeurIPS 2020

תחומי מאמר:

- תמצאות אבסטרקטיבי (abstractive summarization) של טקסטים
- למידה באמצעות חיזוקים (RL - reinforcement learning)

כלים מתמטיים, מושגים וסימונים:

- טרנספורמרים
- פונקציית מטרה סרגיאט (surrogate objective - F_sur)
- proximal policy optimization (PPO)
- שיטות אזור אימון (trust region TR)
- פונקציית גמול (reward function)
- מרחק KL
- מבחן ROUGE

תמצית מאמר:

המאמר מציע שיטה לשימוש יעיל בתיאוג אנושי של>Data עבור שימושות תמצאות אבסטרקטיבי של טקסטים. תמצאות אבסטרקטיבי של טקסט/פואט הינו סיכון קצר (עד 48 טוקנים במאמר זה) של עלילתו, שלא בניי מהמשפטים מהtekst המקורי (אם תנאי זה לא מתקיים המשימה נקראת תמצאות אקסטרקטיבי).

כמו שאתם אולי ידועים, רוב המודלים לתמצאות אבסטרקטיבי היום מאמנים לחקות את התמציאות שנכתבו ע"י בני אדם וביצועיהם נבדדים לרוב בשיטות השוואה בין קטעי טקסט כמו ROUGE. המאמר מצין כי הן שיטות האימון והן מטיקה למדידת ביצועים אלו לא מספקות אינדיקציה מסוימת טוביה על איות התמציאות, המהווה המדריך הטבעי החשוב ביותר לביצוע מודלים לתמצאות אבסטרקטיבי.

בעקבות זאת בשנים האחרונות נעשו מאמצים לשלב משוב (פידבק) אנושי בתהילך אימון של מודלים לתמציאות אבסטרקטיבי. גישות אלו מבוססות לרוב על דירוג של תמציאות, שגונרטו באמצעות המודל, ע"י בני אדם. הבעיה העיקרית עם גישה זו היא סקלביות - רוב המודלים המודרניים לתמציאות מכילים מיליארדים של פרמטרים ונדרשים נתונים אטומטיים מאוד גדולים בשבייל לאמן אותם בצורה מסוימת טוביה.

המאמר הנסקר מציע גישה יותר יעילה לניצול של פידבק אנושי על תמציאות. השיטה המוצעת מקטינה באופן נicer את כמות התמציאות שצריך לתייג. בגודול המאמר מציע לדגום זוגות של תמציאות של אותו טקסט/פוסט מכמה מודלים מאומנים לתמציאות. לאחר מכן המתאים (בני אדם) מוחליטים מה התמציאות היוצרת טוביה מכל זוג של תמציאות - ככה המחברים בונים את הדאטאסת שלהם. לאחר מכן הם מאומנים מודל (נסמן אותו ב-abs_M), המשערת את איות התמציאות על סמך תוצאותיו (כלל שהטמציאות יותר טוביה, היא מקבלת ציון גובה יותר). בשלב האחרון מרצים שיטת מעולם של למידת אמצעות חיזוקים, הנקראת PPO, כאשר המטרה הינה לאמן מודל הבונה תמציאות אבסטרקטיביות תוך כדי מקסום הציון נתן ע"י abs_M.

הסבר של רעיונות בסיסיים:

כמו שכבר אמרנו בתהילך המוצע במאמר מכל 3 שלבים עיקריים:

- **בנייה דאטאסת pair D:**

שלב זה הוא היחיד שבו נרדשת התערבות אנושית. נתונים לכל מתיאג טקסט ושתי תמציאותיו שנדרגו מאחד המודלים שאומנו לגנרטת תמציאות. המתיאג צריך לסמן את התמציאות הטובה מבין השתיים. "טוב" התמציאות מוגדרת לפי שני הקритריונים הבאים: התמציאות צריכה להוות סיכון טוב של עלייה הטקסט ועליה להיות מסויק קצרה (עד 48 טוקנים). נציין כי המתיאגים לא נתונים שום ציון רק לתמציאות, רק נתונים ל"יביל 0" לתמציאות פחות טוביה ול"יביל 1" לתמציאות טוביה יותר מהשתיים.

- **אימון מודל המשערת את איות התמציאות score_M (בהינתן הטקסט כМОבון)**

כאן לוקחים טקסט ושתי תמציאותיו ומעבירים אותם למודל (רשת נירונים כМОבון), המשערת את "איכוטן". המודל פולט שני ציונים (אחד לכל תמציאות) כאשר פונקציית לוס מונסה למקסם את הפרש בין הציונים של תמציאות טוביה יותר לבין הפחות טוביה מהזוג. בדרך זו התמציאות היוצרת איכוטיות יקבלו ציונים גבוהים ואילו הפחות טובים יקבלו ציונים נמוכים יותר. לאחר אימון המודל מkapיאים את המשקליו ועוברים לשלב הבא. שימושו לב בששלב זה לא מאמנים שום מודל לבניית תמציאות - רק את המודל שמחשב את ציון התמציאות בהינתן טקסט. פונקציית לוס כאן היא לוגיריתם של הסיגמוואיד של הפרש הציונים.

- **אימון מודל לתמציאות אבסטרקטיבי על סמך score_M.**

מריצים אלגוריתם PPO מעולם למידת באצעות חיזוקים בשבייל לאימון מודל לתמציאות אבסטרקטיבי score_M, אשר פונקציית גמול rew_F היא הציון שניתן לתמציאות ע"י המודל score_M. זאת אומرت מנסים לאמן מודל לגנרטת תמציאות בעלי ציון גובה. המטרה כאן היא לאמן המודל score_M (שהוא בעצם הפוליסי במקרה זהה) כך שהוא ימקסם את rew_F. בעצם המאמר לוקח מודל מאומן לתמציאות וועושים לו כיול בדרך זו.

אם נסתכל על הנוסחה של פונקציית גמול rew_F, נגלה כי יש בה עוד איבר, המכיל מרחק KL (עם מינוס) בין התפלגות הפלטים (מוחנה בטקסט הקלט) לבין המודל הנלמד באמצעות PPO לבין המודל

הנלמד בתהליך אימון רגיל (ללא שימוש בליבלים על זוגות תמציות - נקרא לו מודל בייסלי). יש להזrina מטרות: המטרה הראשונה היא למנוע "מודולפוף" של מודל מובסס PPO. המטרה השנייה היא מניעת "התרכחות יתר" של מודל PPO מהמודול הבייסלי. כאן יש הנחה סמייאתית המודול הוא לא צזה גרוועציר לשפר אותו רק "בקטנה" בשבייל להציג ביציעים טובים טובים. צריך לציין שהמחברים השתמשו בארכיטקטורה של הטרנספורמר (בסגןן 3-GPT) לגינרט תמציות בכל המודלים שלהם.

הסבר על מושגים חשובים במאמר:

עקרונות של אלגוריתם PPO: אלגוריתם זה שיר' למשחת שיטות policy gradient policy שהיא בעצם הכללה של שיטת TR הקלאסית. Trust Region (TR) מנסה לאמן מודל פונקציית פוליסי ($\pi_{\text{pol}}(s, a)$) שבמוקם למוקם פונקציית גמול $\pi_{\text{old}}(s, a)$, ממקסם פונקציית גמול חלופית (surrogate) $\pi_{\text{sur}}(s, a)$. פונקציה חלופית זו מנסה לשפר את הפוליסי π_{pol} על מרחוק המצביעים של פונקציית היתרון π_{adv} המוכפלת ביחס של π_{old} החדש. בדרך זו π_{old} החדש לומדת לתת הסתברויות גבוהות לצביעים שבHAM פונקציית היתרון מקבלת ערכים גבוהים. דרך אגב השם של השיטה (אזור אימון) נובע מהעובדת שבעית אופטימיזציה זו פורטטים תחת אילוץ שככל עדכון של π_{old} מרחוק KL בין π_{old} החדש לשינה חסום ע"י קבוע קטן. אילוץ זה נדרש בשבייל לא לתת ל π_{old} "להתפרק" כי השונות בבאטיצים עלולה להיות גבוהה. קיימים כמה צורות של פונקציית גמול חלופית π_{sur} שאחת מהן, למשל, משנה את ערך המקסימלי של מרחוק KL כפונקציה של מוצע של מרחוק KL בין π_{old} החדש לשינה בכמה באטיצים אחרים.

- PPO מאמצת גישה דומה לבניית פונקציית מטרה של מוסיפה אליה שתי תוספות:
- מוסיפה לפונקציית מטרה את השגיאה הריבועית המוגיעה של שעורר פונקציית ערך (value function) על הבاطץ.
 - מנסה לשפר את יכולת גילוי (exploration) של π_{old} על מוקסם של האנטרופיה שלה.
- נציין כי המאמר בחר להשתמש בשתי רשומות שונות לשערוך של π_{old} ושל פונקציית ערך.

מדד (מבחן) ROUGE: משווה בין שני קטעי טקסט ע"י השוואת סטטיסטיות על ח-גרמים (n-gram) בין הקטעים.

הישגי מאמר:

המאמר משווה את יכולות התמצאות של המודלים שאומנו באמצעות הגישה המוצעת, מול המודלים שאומנו ללא התערבות אנושית כאשר מספר פרמטרים במודלים שווה (כאן הם לוקחים בחשבון גם את המודול מהשלב השני). המאמר מראה כי עברו אותו מספר פרמטרים המודול שלהם מוציאה תמציאות יותר איותיות (ההשוואה מתבצעת ע"י אדם ש শ্রেণী | איזה מהתמציאות יותר טובה). בנוסף הם מראים שיכולה ההכללה של השיטה המוצעת יותר טובה מאשר מודל SOTA (מאנים על דאטאסטט מודמיין טקסטואלי מסויים ומריצים בדומיין אחר). המachers גם משווים את יכולות התמצאית בכמה פרמטרים שונים כמו Kohonen, דיק וESIS וגם כאן הם משווים את המתחרים מאחור (לאוטו מספר של פרמטרים).

נקודה מעניינת: המאמר מצין (בצירוף דוגמאות) כי יכולות התמצאית מגיע למקסימום כאשר מאנים את המודול abs_{abs} מסווק זמן (היא לא עולה אם מזריםים אליה דוגמאות נוספות וממשיכים לאמן) ולא מספקים הסבר לכך. אני חשב שזה נובע מה הצורך של פונקציית המטרה המוצעת, המשלבת מוקסם של ציון התמצאית תוך כדי שימוש מרחק KL קטן בין המפלגות התמצאית המגנרט ע"י לבין ההתפלגות המגנרט ע"י מודול ללא התערבות אנושית. אני חשב שזה גורם ל PPO "ליצור דוגמאות אדוורסריות" קרי לא שינוי משמעותית בהתפלגות הפלטאגרים לשינוי גדול בציון שלו.

ג.ב.

מאמר מראהים עם רעיון מקורי המשלב טכניקות מלמדית באמצעות חיזוקים שאנו לא מראים לראות במאמרי NLP. השיטה שלהם מעלה את יכולות הניצול של הפידבק האנושי אך עדין יקרה מדי (ל AI OPEN אין בעיות תקציביות) כדי לבנות מודלים לתמצאות אבסטראקטיבי בדומיינים אחרים.

Review 54: Robust Optimal Transport with Applications in Generative Modeling and Domain Adaptation

פינט הsofar:

המלצת קריאה ממייק: מומלץ למביני עניין בטכניקות מורכבות ל-domain adaptation.

bahiorot כתיבה: ביןנית

רמת היכרות עם כלים מתמטיים וטכניקות של DL/ML הנדרשים להבנת מאמר: הבנה عمוקה בתכונות של מרחקים שונים בין מידות הסתברות והבנה טוביה בעיות אופטימיזציה עם אילוצים. הבנה בטרנספורט אופטימי רצiosa גם כן.

ישומים פרקטיים אפשריים: ניתן להשתמש בגישה זו לאימון שלGANים כאשר סט האימון חזוד להציג דוגמאות זרות וגם כן למשימותUDA.

פרטי מאמר:

lienק למאמר: [זמן להורדה](#).

lienק לקוד: [זמן כאן](#)

הורסם בתאריך: 12.10.20, בארכיב.

הציג בכנס: NeurIPS 2020

תחום מאמר:

- מרחק בין נתונים עם אווטליירים (outliers)
- מודלים גנרטיביים (GANs)
- אדפטציית דומיינים בלתי מונחת (unsupervised domain adaptation - UDA)

כלים מתמטיים, מושגים וסימונים:

- טרנספורט אופטימלי (OT)
- טרנספורט אופטימלי רובוטי (ROT)
- טרנספורט אופטימלי בלתיamazon (UOT)
- מרחק וירשטיין (WD), מרחק f ומרחק χ_2 בין מידות הסתברות ([f-divergence](#))
- בעיות אופטימיזציה מינימקס (minimax problems)
- פונקציות ליפשיץ עם מקדם 1 (Lip-1)
- דוגמאות לא טיפוסיות או אוטילירם (OL)

תמצית מאמר:

המאמר הנסקר מציע שיטה לחישוב מרחק בין דאטסהטים, הרובוטי לדוגמאות לא טיפוסיות (OL, outliers). למעשה המרחק המוצע מוגדר עבור כל שתי מידות הסתברות והמרחק בין דאטסהטים הוא המקה הפרטני שלו. מרחק זה נקרא טרנספורט אופטימלי רובוטי (ROT - Robust Optimal Transport) והוא מבוסס על מרחק OT הסטנדרטי ומנסה להתגבר על רגשותו לדוגמאות OL. המאמר דין ברובו במקה הפורטי של OT שהוא מרחק וירשטיין (WD - Wasserstein Distance) כרך שאטמךד רק במרחק וירשטיין הרובוטי (RWD) בהמשך הסקירה. רגשות של מרחק OT לדוגמאות OL ניתן לנוכח באופן הבא: בהינתן שני דאטסהטים עם WD די נורא, החלפתו של חלק מאד קטן של דוגמאות באחד דאטסהטים בדוגמאות OL עלולה להוביל לעלייה בלתי פרופורציונלית ב-WD ביניהם. לטענת המאמר מרבית הדאטסהטים הגדולים מכילים דוגמאות OL, ושימוש במרחק ביניהם שרגיש לדוגמאות אלו, עלול להוביל לתוצאות ירודות במשימות שונות. למשל אימון של GAN עם מטריקת מרחק ציז'ו (כמו וירשטיין גאן - WGAN) עלול להוביל לכך ש-WGAN יגנרט "ערובים" בין הדוגמאות הרגילותות לבין דוגמאות OL.

רעיון בסיסי:

אחת הדרכים להתמודד עם סוגיה זו היא משקל דוגמאות OL במטרה למזער את השפעתן על המרחק. טרנספורט אופטימלי בלתיamazon (UOT) משתמש בReLU הינה ומציע לשערק את המרחק בין התפלגיות P_1 ו- P_2 ע"י המרחק בין שתי התפלגיות קרובות אליהן, Q_1 ו- Q_2 בהתאם, ע"י הוספה של שני איברי רגולריזציה המכילים את סכום המרחקים $\text{Div}(Q_1, P_1) + \text{Div}(Q_2, P_2)$. המרחק $\text{Div}(P, Q) = \text{Div}(P, Q_1) + \text{Div}(Q_1, Q_2) + \text{Div}(Q_2, Q)$ מוגדר כמרחק f -divergence. לעומת פונקציה f נתונה. הביעתיות בגישה זהה נובעת בהיבט המימושי שלה. בדרך כלל לא פתרים את בעיית הטרנספורט האופטימלי בצורה ישירה אלא פותרים את הבעיה הדואלית שלה (הידועה כצורה של קנטרוביץ'-רוביינשטיין). להבדיל מביעית טרנספורט אופטימלי הסטנדרטית, הצורה הדואלית של OT מכילה שתי פונקציות שאווןן צריך לאפותם בו זמנית (כאשר הן תלויות אחת בשניה בדרך די מורכבת) שמקשה מאוד על יישומו לבעיות פרקטיות כמו אימון של GAN.

בשביל להתגבר על קושי זה ולשמור את הרובוטיות של המרחק לגבי דוגמאות OL, המאמר מציע לשנות את ניסוח בעיית אופטימיזציה של UOT באופן הבא: במקום לאפטם על כל התפלגיות ה "בערך שות" ל- P_1 ו- P_2 , הם "מגבילים" (מלמעלה) את המרחקים הללו ע"י קבועים ρ_1 ו- ρ_2 . זה מבון הופך את מרחק ROT המוצע במאמר להיות תלוי בפער ישר- $\rho_1 - \rho_2$ אבל הבעיה הדואלית נהיית יותר פשוטה ותלויה רק בפונקציה אחת (שהיא פונקציה ליפשיץ מסדר 1). מצד שני זה מօסיף אילוץ לבעיית אופטימיזציה הדואלית אך המאמר מוכיח שעדיין ניתן לפתור אותה בדרך יחסית נוכה.

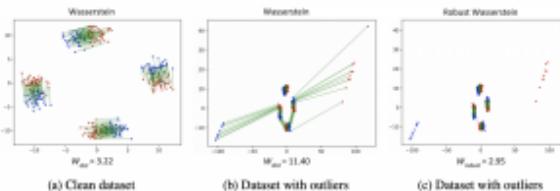


Figure 1: Visualizing couplings of Wasserstein computation between two distributions shown in red and blue. In (a), we show the couplings when no outliers are present. In (b), we show the couplings when 5% outliers are added to the data. The Wasserstein distance increases significantly indicating high sensitivity to outliers. In (c), we show the couplings produced by the Robust Wasserstein measure. Our formulation effectively ignores the outliers yielding a Wasserstein estimate that closely approximates the true Wasserstein distance.

תקציר מאמר:

בזכר קודם כל מה זה מרחק OT והמקרה הפרטני שלו מרחק וורשטיין WD.

טרנספורט אופטימלי OT:

OT הינו מרחק בין שתי מידות הסתברות P_1 ו- P_2 , המוגדרות על אותו מרחב X , עבור פונקציית מחיר Ai שלילית y_1, y_2 (y₁, y₂) נתונה. OT מודד עד כמה מידות הסתברות "קרובות" (כמו מרחק KL או JS). מקרה פרטי של OT שבו פונקציית מחיר הינה מרחק d , נקרא מרחק d , נקרא מרחק wd מסדר d. כאשר $d=1$ נקרא wd מרחוק earth mover.

איך wd בעצם?

בנוסחה עבור wd בין P_1 ו- P_2 מופיע מינימום מעל כל מידות הסתברות על מרחב המכפלה של X עם עצמו, כאשר הפונקציות השוליות שלהן הן מידות הסתברות P_1 ו- P_2 . וחתת סימן האינטגרל יש את המרחוק בין הנקודות. לפשרות בואו ניקח $d = k$. בנוסף נניח שמרחב X הוא חד מימדי (R). למה זה בעצם נקרא wd earth mover? למעשה wd מגדיר כמה "מסה" אנו צריכים להעביר בשיבול להפוך את המידה הסתברות P_2 ל- P_1 כאשר המחיר העברת הנקודה x מהתומך P_2 לנקודה y מהתומך של P_1 הינה $|y-x|$.

למה פעולה מינימום מופיעה בנוסחה עבור wd, אתם שואלים? אפשר "להפוך את P_1 ל- P_2 במספר דרכים ואנחנו רוצים את הדרך הכי קצרה (מבחינת "הمسה המועברת").

ולמה מופיעה בנוסחה מידת הסתברות M על מרחב המכפלה של X עם עצמו? פונקציה ב-(y,x) זו מגדירה איזה " חלק" מהמסה ההסתברותית בנקודה x אנו מעבירים לנקודה y. נניח שלנקודה x הסתברות 0.5, אנו מעבירים שלישי ממנה לנקודה 1 y ושני שלישי הנורטורים לנקודה 2 y. במקרה זה $0.5 * 1/3 = 0.17$ $= 0.5 * 2/3 = 0.33$ $= 0.5 * 1/3 = 0.17$. התנאי שהפונקציות השוליות של M צריכה להיות שוות P_1 ו- P_2 נחות, כי אנו רוצים להעביר את כל המסעה מכל הנקודה של P_1 לנקודות של P_2 בלי לאבד (או להרוויח) מסה נוספת. להבדיל כמעט כל מרחק בין מידות הסתברות wd לוקח בחשבון של התכונות של הקבוצות שעליהן מידות אלו מוגדרות בצורה מפורשת ע"י התחשבות במרקם בין הנקודות שלהם. ולבסוף הצורה הדואלית של wd היא בעצם בעיית אופטימיזציה המנסה למקסם הפרש התוחלות של פונקציית ϕ תחת P_1 ו- P_2 מעל מרחב של כל פונקציות ϕ מסדר 1.

עכשו בואו נסביר איך ניתן להגיד wd על נתונים:

איך מגדירים wd בין נתונים?

עבור שני דאטהסיטים בגודל סופי ניתן להגדיר את מידות ההסתברות עליהם בסכום של פונקציות דלתא על הנקודות של דאטהסיט, כאשר ההסתברות של כל נקודה הינה שווה. המרחק בין כל הנקודות בדאטהסיטים ניתן ע"י מטריצה ואז בעיית אופטימיזציה הופכת לבעית תכנות לינארית (המידה על מרחב המכפלה שעליה מביצעים את האופטימיזציה ניתנת לתיאור ע"י מטריצה גם כן).

הדבר האחרון שנזכר לנו זה להבין איך WD הרובוטי (RWD) מוגדר על דאטהסיטים:

איך מוגדרים RWD בין דאטהסיטים?

קודם כל נזכיר כי כל אחת פונקציות התפלגות (מידות הסתברות) Q_1 ו- Q_2 קרובות ל- P_1 ו- P_2 בהתאם (ראה הסבר בפרק "רעיון בסיסי") ניתן להגדיר בתור משקל של הסתברויות של דוגמאות (כਮון שגם $b_1 P_1$ וגם $b_2 P_2$ לכל דוגמא של אותה הסתברות) בשני הדאטהסיטים כאשר סכום המשקלים בדאטסיט הוא 1 (אחרת Q_1 לא תהווה מידת הסתברות). ניתן לראות כי בעית אופטימיזציה שאנו פותרים כוללת שני סטים של משקלים המסתכניםים ל-1 (ועל כל פונקציות -1 -kilo). להבדיל מ-WD מתווספת כאן המגבלה על המרחקים בין ההתפלגות של הסטים הממושקלים למקוריים (צריכים להיות קטנים מ- $1 - \rho$ ו- $2 - \rho$). המאמר מראה כי תנאים אלו ניתן לתרגם למרחק χ^2 בין ההתפלגותים הממושקלות Q_1 ו- Q_2 למוקוריים P_1 ו- P_2 על הדאטהסיטים. בעיה זו למשה הינה [תכנות קוני מסדר שני](#) ויש דרכים יעילות לפטור אותה. עבור דאטהסיטים גדולים לפתורן מהברור המאמר עשו רפרמטריזציה של המשקלים ע"ר רשותנו נירונים כאשר הקטל לרשותן אלו הוא דוגמאות מהדאטהסיטים.

הישגיו מאמר:

המאמר השתמש ב-RWD כדי לבנות GAN עם המציג RWD בין הגנרטור לבין דאטהסיט האימון. המחברים הראו כי עבור דאטהסיטים המכילים דוגמאות OL (או אלו שהם יוצרים באמצעות "לכלור" דאטהסיטים "נקויים") באחד מוסיפים של תמונות מדאטהסיטים אחרים) התמונה שוגנרטו עם RWDGAN נראות יותר "נקיות" מבחינה ויזואלית אפילו עבור אחוז OL יחסית גבוהים. מעניין כי כאשר מאמנים את RWDGAN על דאטהסיטים נקיים (עם $1 - \rho$ ו- $2 - \rho$ מסוימים) יש פה שאלה של איך לכייל אותם) אז IS ו- FID של התמונות המוגנרטות איתו כמעט ולא השתנה ויחסית לאימון עם WD רגיל. ההשוואות נעשו כאן עבור ורסיטיון גאן עם 3 ארכיטקטורות שונות.

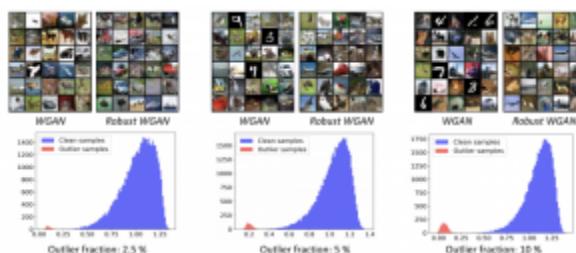


Figure 3: Visualizing samples and weight histograms. In the top panel, we show samples generated by WGAN and robust WGAN trained on CIFAR-10 dataset corrupted with MNIST samples as outliers. WGAN fits both CIFAR and MNIST samples, while robust WGAN ignores the outliers. In the bottom panel, we visualize the weights (output of the $W(\cdot)$ function) for in-distribution and outlier samples. Outlier samples are assigned low weights while in-distribution samples get large weights.

תופעה מעניינת של RWDGAN: משקל אופטימלי של דוגמא נתונה למשה משקף את "רמת הקושי" של הגנרטור לגנרטו אותה (כלומר עד כמה דיסקרמיןטור הצלח "לפוץ אותה"). אתם שואלים למה בעצם? אם משקל של דוגמא נמוך, זה אומר שהגנרטור "החליט להנmir בחשיבותה ולהקטין את השפעתה ללו"ס" מהסיבה שהוא חושב שהדוגמא הזאת OL. ד"א המאמר מראה שבdatashitim "מלוכלים" עם דוגמאות מדאטהסיטים אחרים המשקלים של הדוגמאות "הזרות" יצאו נמכות משמעותית מהרגילות.

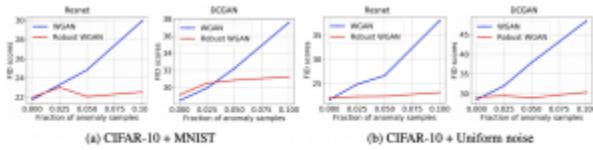


Figure 2: FID scores of GAN models trained on CIFAR-10 corrupted with outlier noise. In (a), samples from MNIST dataset are used as the outliers, while in (b), uniform noise is used. FID scores of WGAN increase with the increase in outlier fraction, while robust WGAN maintains FID scores.

בנוסף המאמר הראה כי שימוש ב-RWD עברו שימושותUDA מיפור באופן ניכר את ביצועי דיקן עבור 3 ארכיטקטורות רשתות שונות (עבור DATA17).
3. ג.ב.

המאמר עם רעיון ד' מעניין, מכיל גם הוכחות ריגורוזיות המסבירות למה הגישה המוצעת עובדת. מה שמטריד אותו טיפה עם RWD זו הבחירה של פרמטר softmax . המאמר מוכיח כי עם אחוז דוגמאות OL ידוע אז קיימים ביטוי לערך softmax אופטימלי. ברוב המקרים זה לא המצב ובחירה של softmax עלולה להיות לא טרייאלית.

שנקרא:

Review 55: InfoBERT: Improving Robustness of Language Models from an Information Theoretic Perspective

פינית הסוקרי:

המלצת קרייה ממייק: חובה בהחלט לאוהבי נושא של אימון אדוורסרי ותורת המידע. לאחרים מומלץ

מאוד

בהירות כתיבה: בינוי פלוס

רמת היכרות עם כלים מתמטיים וטכניקות של DL/ML הנדרשים להבנת מאמר: היכרות עם עקרונות של התקיפות אדוורסריות לרשותות נוירונים (בדגש על NLP), הבנה טוביה במושגי יסוד של תורת המידע כמו מידע הדדי של משתנים אקראיים.

ישומים פרקטיים אפשריים: אימון מודלי NLP, עמידים להתקפות אדוורסריות.

פרטי מאמר:

lienck למאמר: [זמן להורדה](#).

לינק לקובץ: [רשמי, לא רשמי](#)

פורסם בתאריך: 22.03.21, בארכ'יב.

הוגג בכנסו: ICLR 2021

תחום מאמר:

- טרנספורמרים, BERT
- אימון אדוורסרי - adversarial training
- למידת ייצוג - representation learning

כליים ומושגים מתמטיים במאמר:

- צואר בקבוק מידע (information bottleneck) ברשותות נירוניים
- מידע הדדי (mutual information)
- InfoNCE (noise contrastive estimation)

תמצית מאמר:

המאמר הנסקר מציע שיטה להתמודדות עם התקפות אדוורסריות כנגד מודלי שפה גדולים בסגנון BERT (עם קלט טקסטואלי). הגישה המוצעת מבוססת על העיקרון של צואר בקבוק מידע (information bottleneck) עבור רשותות נירוניים. עקרון זה מגדיר את מטרת האימון של רשת נירונים כמייקסום של פונקציית מטרת lb_\perp כאשר lb_\perp היא הפרש בין שני איברים (כל אחד מהם הינו מידע הדדי). האיבר הראשון משערך יכולת חיזוי של רשת והאיבר השני מודד את מידת דחיסת קלט ע"י רשת.

המאמר מציע להוסיף ל- lb_\perp איבר נוסף, המזקסם את המידע היחידי של ייצוג הקלט (משפט או מקטע של טקסט עבור מודלי שפה) לבין ייצוגי טוקנים, שנקבעים במאמר localized anchored tokens. טוקנים localized anchored הינם טוקנים רובוטיים (חסינים) להתקפות אדוורסריות וגם מועילים למשימת downstream.

הטענה המרכזית של המאמר (מוחחת בחלוקת תיאורטיב ובחילקה אמפירית) שאימון מודל שפה עם פונקציית המטרה המוצעת משפר את הרובוטיות (חסינים) של הרשת נגד דוגמאות (התקפות) אדוורסריות. מעניין כי המאמר מראה (אמפירית) שטענה זו נכונה גם עבור אימון רשת על דאטאטטים רגילים ללא דוגמאות אדוורסריות וגם באימון על דאטאטטים, המכילים דוגמאות כאלה.

רעיון בסיסי:

מחברי המאמר טוענים (ומוכחים ריגורוזית) שאימון רשת נירונייםقلלית עם פונקציית המטרה המוצעת מקטין את ההפרש בין:

- מידע הדדי, המסומן ב-(Y,T), של ייצוג קלט נקי (לא אדוורסרי) והחיזוי של רשות עברו קלט זה (לייבל)
- מידע הדדי בין ייצוג קלט מורעש (אדוורסרי) לבין אותו החיזוי של רשות, המסומן ב-(Y,T).

בנוסף אימון עם פונקציות מטרה כזו ממקסם את מידע הדדי בין ייצוג הקלט לבין הליבלים הנחוצים באמצעות רשות, המתורגם לbijoux מודל במשימת downstream.

למה זה טוב, אתם שואלים? שימו לב שבוספו של דבר המטרה של האימון האדוורסרי הינה הפיכת הרשות לעמידה נגד דוגמאות אדוורסריות. ככלומר חיזוי של רשות לא אמרו להשתנות כאשר הופכים דוגמא רגילה לדוגמא אדוורסית (צריך לזכור כי בדרך כלל משנים דוגמא בכוונה מינורית כדי להפוך אותה לאדוורסית). עקב העובדה ש-(Y,T) ו- (Y,X) מהווים מدد לביצועי רשות עם קלט אדוורסרי ורגיל בהתאם, מזעור ההפרש ביניהם מתורגם (לטענת המאמר) לbijoux טוביים יותר של מודל בתרכיש אדוורסרי.

תקציר מאמר:

בשביל להבין את רעיון המאמר אנו צריכים להבין מה זה בעצם עיקרונו צואר בקבוק מידע ברשותות נירוניים:

עיקרונו צואר בקבוק מידע ברשותות נירוניים:

עיקרונו צואר בקבוק מידע (שהומצא ע"י פרופ' תשבי ב-2015) מגדיר את מטרת למידה עמוקה (כלומר אימון של רשות נירוניים) כטרייד-אוף בין דרישת מידע ע"י רשות (בנייה ייצוג דחוס של קלט) לבין יכולת החיזוי שלה. עיקרונו זה מתורגם למיקסום מידע הדדי בין ייצוג קלט T לבין חיזוי של רשות Y, המסומן (Y,T). ובאותו זמן למינימיזציה של מידע הדדי בין קלט X לייצוג עצמו, המסומן כ- (T,X). שימו לב כי (Y,T) מהווה אינדיקציה לביצועי רשות על סט האימון. לעומת זאת (T,X) אפשר לפרש כאיבר רגולרייזציה למזעור אוברפיטינג (overfitting).

המאמר הנסקר מציע לאמן מודל שפה על משימת downstream עם פונקציית מטרה שבilibה עיקרונו צואר הבקבוק של מידע. כמו שכבר ראינו קודם הפונקציה המוצעת מכילה את מידע הדדי בין ייצוג של קלט T לבין קלט X (שבמקרה שלנו הינו משפט). T מכיל את האמבדינגים ("יצוגים") של טוקנים, המרכיבים משפט X, כאשר מימד של ייצוג של טוקן i_T הוא 768 (עבור Base BERT). המימד הגובה של T אינו מאפשר לחשב/לשער את (T,X) בכוונה ישירה. המאמר מציע (ומוכיח ריגורוזית) שניתן לחסום (T,X) מלמטה ע"י סכום של (i_T, X) המוכפל במספר הטוקנים במשפט. שימוש ביחסים זה הופך את בעיית אופטימיזציה זו לקללה יותר מבחינה חישובית.

דוגמה אדוורסרית בעולם NLP:

בשביל להמשיך את ניתוח המאמר בואו נבין מה זה דוגמא אדוורסרית בדומיין של NLP. נזכיר שדוגמא אדוורסית נוצרת באמצעות הוספת רעש קטן לדוגמא רגילה כדי לעוות את הליביל הנחזה עבורה באמצעות הרשות. בדומיין טקסטואלי משפט אדוורסרי נוצר ע"י שינוי של משפט רגיל השומר מרוחקים בין האמבדינגים של המילים במשפט המקורי לבין המילים של " המשפט האדוורסרי" קטנים. שינוי זה לא גורם לשינוי הליביל של המשפט ("א מתיאג אנושי היה מעניק למשפט את אותו לייבל כמו למשפט המקורי") אך הוא כן "מבלב את הרשות" שמשנה את החיזוי שלה עבור המשפט המורעש (האדוורסרי).

כבר אמרנו כי המאמר הנסקר מציע להויסיף לפונקציית המטרה המקורית d_L איבר רגולרייזציה נוסף, הממסם את סכום של המידעים ההדדיים של ייצוג משפט Z והיצוגים של הטוקנים הנקראים במאמר (local anchored) (LA).

טוקני LA:

יצוגים של טוקנים LA הם בעלי שתי התכונות הבאות:

- רובוטיים (חסינים) בתרחישים אדוורסריים.
- מכילים מידע מועיל למשימת downstream.

המאמר מציע לאתר טוקנים בעלי תכונות אלו באמצעות איתור של טוקנים שדווקא לא מקיימים את הדרישות הללו (!!). כדי לאთר את טוקנים לא חסינים נגד התקפות, המאמר מציע "לבצע" התקפות אדוורסריות על הטוקנים. המטרה כאן היא לזרוח טוקנים שניינו קטן ביצוגם מביא לעלייה משמעותית בלוס של ממשמה downstream. טוקנים כאלה מהווים מועדים נוחים לבניה של דוגמא אדוורסרית על גיביהם. מצד שני יש לנו טוקנים כמו stopwords או סימני פונקטואציה שאיפלו שניינו גדול באמבידיג שלהם לא גורם לעלייה גדולה בלוס של המשימה. עם זאת טוקנים אלו כללים לא מועילים למשימה. בהקשר זה המאמר מציע לאתר טוקנים שניינו מותן בהם גורם לשינוי מותן בלוס עבור משימת downstream ולנצל אותם כ"עוגני אמבדינג של המשפט".

מכיוון שאנו רוצים לנצל כמה שייתר את המידע מטוקני LA בשביל לבנות ייצוג משפט עמיד נגד דוגמאות אדוורסריות. זו הסיבה שמוסיפים סכום של כל המידעים ההדדיים בין ייצוג המשפט Z וטוקני LA לפונקציית מטרה.

אין מושרכים את פונקציית המטרה בפועל:

از הכל טוב ויפה אבל נשאלת השאלה איך אנחנו נאמן רשות פונקציית מטרה שלה כוללת כל מיני מידעים הקיימים בין וקטוריים אקראים שונים? הרי ידוע שישירוך של מידע הדדי הינו untractable ובדרך כלל משתמשים בחסמים בשביל לבנות פונקציית מטרה שהיא יותר נוחה לאימון רשות. במאמר הנסקר נעזרים ב-InfoNCE עם פונקציית מרחק נטוונה d בין הייצוגים ובונים פונקציית מטרה ש"מקרבת" את הייצוגים שאנו רוצים למוקם את המידע הדדי ביניהם (כמו ייצוג המשפט והטוקנים LA), "ומרחיקה" את הייצוגים של טוקנים ומשפטים הנבחרים בצורה רנדומלית. פונקציית מרחק d יכולה להיות מרחק cosine או שטמודלת באמצעות רשת MLP עם שתיים-שלוש שכבות.

Algorithm 1 - Local Anchored Feature Extraction. This algorithm takes in the word local features and returns the index of local anchored features.

- 1: **Input:** Word local features t , upper and lower threshold c_u and c_l
 - 2: $\delta \leftarrow 0$ //Initialize the perturbation vector δ
 - 3: $g(\delta) = \nabla_{\delta} \text{d}_{\text{L}}(q_\phi(t + \delta), y)$ // Perform adversarial attack on the embedding space
 - 4: Sort the magnitude of the gradient of the perturbation vector from $\|g(\delta)_1\|_2, \|g(\delta)_2\|_2, \dots, \|g(\delta)_n\|_2$ into $\|g(\delta)_{z_1}\|_2, \|g(\delta)_{z_2}\|_2, \dots, \|g(\delta)_{z_n}\|_2$ in ascending order, where z_i corresponds to its original index.
 - 5: **Return:** k_1, k_{1+1}, \dots, k_j , where $c_l \leq \frac{1}{n} \leq \frac{j}{n} \leq c_u$.
-

air זה נעשה? בונים מיני-באטץ', המורכב מזוג אחד של "ցוג משפט וטוקן A" ממנה ("ցוגים קרובים"), כאשר שאר הזוגות מורכבים ממשפט וטוקנים שנבחרו רנדומלית. פונקציית מטרה היא ייחוס כאשר המונה מכיל אקספוננט של מרחק בין "ցוגים של "הזוג הקרוב"" והמננה מכיל את סכום אקספוננטים של המרחקים בין כל הזוגות. Ord et al hocih ב-2018 שמייקסום של פונקציה מטרה מצורה זו מגדיל את מידע הדדי בין "ցוגים של זוגות קרובים כולם מושג את המטרה שלהם CAN.

הישגי מאמר:

המאמר מראה את עליונותה של שיטת האימון (כיול) InfoBert בהטמודדת נגד דוגמאות אדוורסריות עבור BERT ו-ROBERTA עבור כמה דאטאסתים אדוורסריים בעלי דרגות קושי שונות. כמו שכברذكرתי גם אימון של InfoBERT על דאטאסט ללא דוגמאות אדוורסריות וגם עם דוגמאות אדוורסריות גורם למודל המאמן להיות יותר עמיד לרעש.

Training	Model	Method	Dev				Test			
			A1	A2	A3	ANLI	A1	A2	A3	ANLI
Standard Training	RoBERTa	Vanilla	49.1	26.5	27.2	33.8	49.2	27.6	24.8	33.2
		InfoBERT	47.8	31.2	31.8	36.6	47.3	31.2	31.1	36.2
	BERT	Vanilla	20.7	26.9	31.2	26.6	21.8	28.3	28.8	26.5
		InfoBERT	26.0	30.1	31.2	29.2	26.4	29.7	29.8	28.7
Adversarial Training	RoBERTa	FreeLB	50.4	28.0	28.5	35.2	48.1	30.4	26.3	34.4
		InfoBERT	48.4	29.3	31.3	36.0	50.0	30.6	29.3	36.2
	BERT	FreeLB	23.0	29.0	32.2	28.3	22.2	28.5	30.8	27.4
		InfoBERT	28.3	30.2	33.8	30.9	25.9	28.1	30.3	28.2

Table 1: Robust accuracy on the ANLI dataset. Models are trained on the benign datasets (MNLI + SNLI) only. 'A1-A3' refers to the rounds with increasing difficulty. 'ANLI' refers to A1+A2+A3.

ג.ב.

מאמר מציע רעיון מאד מעניין המתבסס על עקרון צוואר בקבוק מידע עבור רשותנו נירוניים לשיפור את עמידות רשות נגד התקפות אדוורסריות. אהבתי את הhocihות הריגורזיות ויפות שיש במאמר (במיוחד משפט 3.2).

Review 56: Meta-Learning Requires Meta-Augmentation

פינת הסוקר:

המלצת קריאה ממיק: מומלץ לאוהבי מטה-למידה אך לא חובה

בהירות כתיבה: גבוהה

רמת היכרות עם כלים מתמטיים וטכניקות של DL/ML הנדרשים להבנת מאמר: נדרשת הבנה טובה של מושגי יסוד של תומם מטה-למידה (meta-learning).

ישומים פרקטיים אפשריים: שיפור ביצועים במשימות של מטה-למידה באמצעות אוגמנטציה של לייבלים.

פרטי מאמר:

[לינק למאמר: זמין להורדה.](#)

[לינק לקוד: זמין כאן](#)

פורסם בתאריך: 04.11.21, בארכיב.

הציג בכנס: NeurIPS2020

תחום מאמר:

- שיטות אוגמנטציה למטה-למידה (meta-learning)
- שיטות התמודדות עם אוברפיטינג (overfitting) במטה-למידה

כליים ומושגים מתמטיים במאמר:

- אפייזודה של שימוש למטה-למידה
- למידה N-way, K-shot
- זיכרון (memorization) במשימות למטה-למידה
- אנטרופיה מותנית (conditional entropy) - CE
- אוגמנטציה שומרת CE

תמצית מאמר:

המאמר הנסקר מציע שיטה חדשה לאוגמנטציה שבאה להתמודד עם בעיית אוברפיטינג (overfitting), המתרחשת במשימות למטה-למידה. המאמר מציע לבצע אוגמנטציה פסאודו-אקראית ללייבלים (ולא לדאטה!!) של המשימות של base learner (מודל חיצוני) ואotta אוגמנטציה גם ללייבלים של המשימות של מודל פנימי. בדרך זו מודל פנימי יהיה "ח'יב" לשחרר את האוגמנטציה שהשתמשה בה במודל חיצוני וכבר לא יכול "להתעלם" מהעדכנים שלו שלפענת המאמר מסייע להתגבר על אוברפיטינג במשימות למטה-למידה.

תקציר מאמר:

נתחיל מלהזכיר מה זה בעיית למטה-למידה:

מה זה בעיית למטה-למידה:

כמו שכולנו יודעים בכל בעיית למיטה supervised נתון לנו סט אימון (Y, X), המכיל זוגות של דוגמאות והלייבלים שלהם (תיוגים). המטרה של אימון supervised היא למדל את הפונקציה המפה X ל- Y .

לעומת למיטה supervised בעיית למטה-למידה יש לנו מספר משימות $_T$, כאשר כל משימה מורכבת מסט תומך (support set), המכיל כמה זוגות של דוגמאות ולהלייבלים (s_x, s_y) וסט שאלתה (query) (q_x, q_y), שבייחד בונים אפייזודה. נציין שבדרך כל גם סט תומך וגם סט שאלתה מכילים מספר מאד קטן של דוגמאות. בנוסף נתונים לנו סט אימון למטה (meta train set), המקביל לסט אימון בעיית ML רגילה ומטטה-טסטס

סט (כמו טסט סט ב-ML רגיל), המכילים כמה אפיוזות כל אחד. המטרה של מטה-למידה היא לאמן מודל (הנקרא **base learner** או מודל חיצוני) על הדadata שבסט התומך (s_u, s_x) כאשר הפלט שלו הינו המודל לחיזוי y_u מ- s_x מתוך סט השאלתה. לעומת המטרה של מטה-למידה היא להקנות למודל החיצוני יכולת "ללמד" את המודל הפנימי (learner).

במודלי' מטה-למידה יש שני שלבי אימון: **השלב הפנימי** שבו מודל חיצוני מעדכן את מודל פנימי במטרה לשפר את יכולת החיזוי שלו עבור דוגמאות מסט שאילתה s_x ובמסגרת **השלב החיצוני** מודלים חיצוניים עצמו במטרה לשפר את יכולתו "ללמד" מודל פנימי. יש כמה סוגים של שיטות מטה-למידה ואחת מהנפוצות מהם היא [MAML](#). ב-MAML המודל החיצוני הוא רשות ניירונים שמאמנים אותה בשבייל לעדכן את המשקלים של המודל הפנימי שהוא גם כן רשות ניירונים.

אגומנטציה: כדיוע המטרה העיקרית של אוגומנטציה של דadata במשימות ML היא מניעת אוברפיטיניג ע"י יצירה של דוגמאות נוספות לאמון של מודל. לאור זה נთאר עתה את סוג אוברפיטיניג המתרכחים במשימות מטה-למידה.

סוגי אוברפיטיניג במודלי' מטה-למידה:

יש שני סוגים עיקריים של אוברפיטיניג שעולמים להתרחש במהלך אימון של מודלי' מטה-למידה:

- **זיכרון (memorization)** - מודל פנימי מתעלם מהעדכנים שמודל חיצוני מעביר לו ומשתמש בפועל רק בדוגמאות שלא מטה השאלתה (לא קיימת בעיות ML רגילות). למעשה במקרה הזה לטט תומך אין שום השפעה על חיזוי של מודל פנימי עבור דוגמאות מסט שאילתה.
- **אוברפיטיניג של learner** - מודל חיצוני עושה אוברפיטיניג על סט אימון מטה ואינו מצליח להכפיל למטה-טסט סט (זה הסוג הרגיל של אוברפיטיניג הקורה במשימות ML סטנדרטיות).

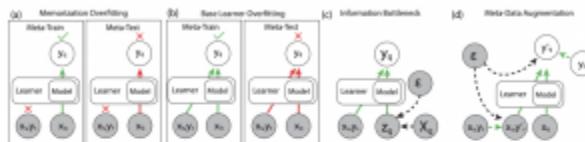


Figure 1: Meta-learning problems provide support inputs (x_s, y_s) to a base learner, which applies an update to a model. Once applied, the model is given query input x_q , and must learn to predict query target y_q . (a) Memorization overfitting occurs when the base learner and (x_s, y_s) does not impact the model's prediction of y_q . (b) Learner overfitting occurs when the model and base learner leverage both (x_s, y_s) and x_q to predict y_q , but fails to generalize to the meta-test set. (c) Yin et al. [37] propose an information bottleneck constraint on the model capacity to reduce memorization overfitting. (d) To tackle both forms of overfitting, we view meta-data augmentation as widening the task distribution, by encoding additional random bits ϵ in (x_s, y_s) that must be decoded by the base learner and model in order to predict a transformed y_q .

בשביל להבין באלו סוגים של משימות מתרכחות אוברפיטיניג מסווג זיכרון אלו צריכים להגדיר את המושג החשוב הבא:

הגדרה: סט משימות נקרא (Mex) mutually exclusive כאשר מודל אחד לא יכול לפתור את כל המשימות ביחד.

למשל אם במשימה אחת מסט המשימות יש תוצאות של סוסים מתיוגות עם לייבל 0 ותוצאות של כלבים המתיוגות עם לייבל 1 ובמשימה השנייה הסוס מקבל לייבל 1 והכלב מקבל לייבל 0, לא קיים מודל שיכל ללמידה את שתי משימות אלו יחד. יש מחקרים שטוענים שסטים משימות Mex הם יותר קלים בתחום מטה-למידה כי מודל פנימי "חייב" לנצל מידע מסט תומך (s_u, s_x) כדי לבצע את המשימה שלא. נראה הסיבה לכך היא שהמודל יתקשה גם "לזכור" את המשימה מהסט התומך, ובאותו זמן למדוד משימה "מנוגדת" למשימה זו מהסט התומך.

הנחה היא שהמודל "יאלץ" ללמידה "פיצרים מועלם" מהדוגמאות מסווג התומך שניצלו לאחר מכן ע"י המודל במהלך אימון על סט השאלתה.

לעומת זאת אם סט המשימות אינו מקיים את תכונת X -Me, אוברפייטינג מסווג זיכרון עלול להתרכש (לטענת המאמר) כי מודל אחד כן יכול ללמידה q_y רק על בסיס q_x בלבד להסתמך על מידע מ- (s_y, s_x) . כאשר זה קורה הביצועים של מודל מטה-למידה טובים על סט אימון מטה וספגים ירידה משמעותית על מטה טסט סט (מטה-הכללה גרוע). הסיבה לכך היא שהמודל החיצוני פשוט "מצרך" את הסט התומך במקום " לניצלו בשבייל ללמידה איך למד את המודל הפנימי".

צריך לציין שרוב המשימות מטה-למידה מסווג *K-shot N-way* (מספר הדוגמאות בכל סט תומך של משימה הינו K ויש בכלל משימה N ליבלים שנדגים באקראי), הסטים של המשימות הינם X -Me כי אנחנו דוגמים אפיוזדות באופן רנדומלי כך שכל קטגוריה מקבלת ליבל שונה בכל משימה. ככלمر במשימה מסוימת החתויל יכול לקבל ליבל 0 כאשר המשימה אחרת הוא יקבל ליבל 1. כאשר המשימות הן מסווג רגסיה העיניים מסתבכים וסטים של משימות מתקשותקיימים את X -Me. כדי להתגבר על בעיות הזיכרון במרקם האלו ניתן להגביל את הזירמה של המידע בין q_x ל- q_y (דרך המידע ההודי שלהם) אבל צריך לעשות את זה בעדינות בשבייל לא להגיע למצב של underfitting.

הסוג השני של אוברפייטינג (learner overfitting) קורה כאשר המודל החיצוני מצלח את הדאטה שלו (s_y, s_x) בשבייל לעזור למשימות של המודל הפנימי בסט אימון מטה אבל אינו מצליח להקליל את זה לאפיוזדות של מטה-טסט סט.

אוקי, אז איך מתמודדים עם אוברפייטינג מהסוג הראשון בלי להגביל את זרימת המידע בין q_x ל- q_y ? בדומה ללמידה הרגילה התשובה היא - אוגמנטציה של דאטה. אבל לא האוגמנטציה הרגילה של הדוגמאות אלא אוגמנטציה של הליבלים. כאמור קוראים לזה מטה-אוגמנטציה.

מטה-אוגמנטציה:

בשביל להבין את הרעיון של מטה-אוגמנטציה בואו קודם בין איזה סוג אוגמנטציה אפשר לעשות לדאטה. קודם כל פועלות אוגמנטציה ניתנת להגדיר בתור מיופיע (Y, X)->(Z, F). אוגמנטציה נקראת שומרת אנטרופיה מותנית (CE preserving) כאשר האנטרופיה של ליבל שעבר אוגמנטציה בהינתן הדוגמא שעברה אוגמנטציה, שווה לאנטרופיה המותנית של הליבל המקורי בהינתן הדוגמא המקורית: $(X|Y)_H = (X|Y)_A$. למשל אוגמנטציה מסווג סיבוב של תמונה תוך הליבל הינה שומרת אנטרופיה מותנית. כמו כן אוגמנטציה נקראת מגדילה אנטרופיה מותנית (CE-increasing) כאשר האנטרופיה המותנית עולה לאחר אוגמנטציה. למשל אם נעשה אוגמנטציה רק ליבל של תמונה נתונה (נוסיף אליו איזה מספר נגיד) אז האנטרופיה המותנית עולה כי באותה תמונה יהיו שני ליבלים שונים.

از המאמר אומר דבר כזה: אנו צריכים אוגמנטציה שתקשר את הזוגות (q_y, q_x) לזוגות (q_y, q_x) כך שמודל פנימי לא יוכל להביא את הlös על סט השאלתה למיניהם ע"י שימוש ב- q_x בלבד אלא "נכricht" אותו "לשתח פועלה" עם s_x . הדרך לעשות זאת היא לעשות אוגמנטציה שהיא CE-increasing למשימות. ככלמר לכל משימה הליבלים s_y ו- q_y ("עוזו") באותה זורה ("עברית" הצפנה) עם אותו מפתח שנבחר רנדומלית או אותה דגימה של רעש). במקרה הזה רשות פנימית יכולה לחזות את q_y המועות מ- q_x רק אם היא הצליחה לענה את מפתח ההצפנה (רעש פסאודו רנדומלי) שהוא יכול ללמוד רק מ- (s_y, s_x) המוצפן.

אינטואיציה לשיטה המוצעת:

אם ניקח משימה מסוימת (אפיוזדה) וניצור סט מספיק גדול של משימות מאוגментות עם אותו מקור של רעש, אז האנטרופיה המותנית של המשימה המוצפנת של המודל הפנימי תעלת ב-(Δ)ה. لكن בשיביל לבצע את המשימה המודל הפנימי ח'יב להקטין את האנטרופיה הزادה באזורה באמצעות "המידה" מהמודל החיצוני.

הישגי מאמר:

המאמר מראה שיפור ביצועים במקרים k-shot, N-way של מספר דאטסהטים המקובלים בתחום מטה-למידה. המחברים הצליחו להקטין את השפעה השילית של אובייקטיבינג מסוג זיכרון בתרחישים שבהם סט המשימות אינו MeX. המאמר משתמש ב-MAML לצורך לאמן את המטה-מודול שלהם. צריך לציין שעבור בעיות סיווג k-shot, N-way המחברים יצרו אפיוזדות כך שסט המשימות שלהם הוא Non-MeX (למרות ש- k -shot, N-way k-shots הוא כן MeX).



Figure 3: Non-mutually-exclusive, intrashuffle, and intershuffle. In this example, the dataset has 4 classes, and the model is a 2-way classifier. In non-mutually-exclusive, the model always sees one of 2 tasks. In intrashuffle, the model sees permutations of the classes in the non-mutually-exclusive tasks, which changes class order. In intershuffle, the model sees $4 \times 3 = 12$ tasks.

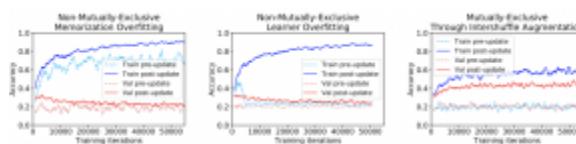


Figure 4: Mini-ImageNet results with MAML. **Left:** In a non-mutually-exclusive setting, this model exhibits memorization overfitting. Train-time performance is high, even before the base learner updates the model based on (x_s, y_s) , indicating the model pays little attention to (x_s, y_s) . The model fails to generalize to the held-out validation set. **Center:** This model exhibits learner overfitting. The gap between train pre-update and train post-update indicates the model does pay attention to (x_s, y_s) , but the entire system overfits and does poorly on the validation set. The only difference between the left and center plots is the random seed. **Right:** With intershuffle augmentation, the gap between train pre-update and train post-update indicates the model pays attention to (x_s, y_s) , and higher train time performance lines up with better validation set performance, indicating less overfitting.

דאטסהטים:

.Omniglot, Mini ImageNet, D'Claw, Pascal3D, Pose Regression

ג.ב.

אהבתי את החשיבה של מחברי המאמר. המאמר קרי, הרעיון מאד אינטואיטיבי ומוסבר בצורה יפה.

Review 57: Geometric Dataset Distances via Optimal Transport

פינט הסוקר:

המלצת קרייה ממייק: חובה למתעניינים בשיטות של domain adaptation.

bahiorot כתיבה: בנויית.

רמת היכרות עם כלים מתמטיים וטכניקות של ML/DL הנדרשים להבנת מאמר: נדרש היכרות בסיסית עם שיטות domain adaptation והבנה טובה בכל מה שקשר לטרנספורט האופטימלי.

ישומים פרקטיים אפשריים: מציאת זוגות של דאטהסטים "נוחים" לביצוע domain adaptation של מודלים ביניהם.

פרטי מאמר:

لينك למאמר: [זמן להורדה](#)

لينك לקוד: לא נמצא בארכיב

פורסם בתאריך: 07.02.20, בארכיב

הוזג בכנס: NeurIPS2020

תחום מאמר:

- אדפטציה בין דומיינים (domain adaptation)
- חיקר של דמיון בין דאטהסטים
- transfer learning

כלים ומושגים מתמטיים במאמר:

- הטרנספורט האופטימלי (optimal transport) •
 - מרחק ורסטיין (WD) •
 - גוסחת רובינסוטיון-קנטורוביץ •
 - שיטת Sinkhorn לחישוב OT •
-

תמצית מאמר:

המאמר הנסקר מציע שיטה למדידת "דמיון" (מרחק) בין דאטהסטים מותאיים. המאמר טוען כי שימוש מרחק המוצע קורלצייה גבוהה למידת הצלחה של domain adaption בין דאטהסטים. למשל נניח שהDataset מודול מאומן על הדאטהסט הראשון וכיילנו אותו (fine-tuning) על הדאטהסט השני. ככל שהמרחק המוצע בין הדאטהסטים קטן יותר, הביצועים של המודול המכוייל על הדאטה מהdomין של הדאטהסט השני, נוטים להיות טובים יותר (לטענת המאמר). בסיום המרחק המוצע הינו אגנוטטי לסוג מודול, לא דורש אימון, לא מחייב שום דמיון בין הליבלים בדאטהסטים ומtabased על הטרנספורט האופטימלי (OT).

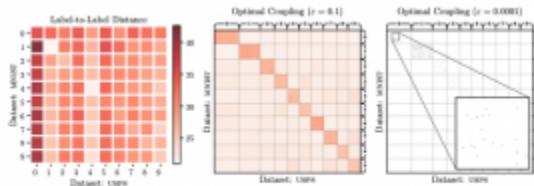


Figure 3. Dataset Distance between MNIST and USPS. Left: The label Wasserstein distances—compared without knowledge of the relations between labels across domains—recover expected relations between classes in the two domains. Centre/Right: The optimal coupling π^* for different regularization levels exhibits a black-diagonal structure, indicating class-coherent matches across domains.

תקציר מאמר:

אני רוצה להתחל עם הסבר קצר על המושגים המתמטיים הנדרשים להבנת המאמר. נתחל מ-OT - המושג המרכזי במאמר.

טרנספורט אופטימלי:

טרנספורט אופטימלי (OT) הינו מרחק המוגדר בין שתי מידות הסטברות P ו- Q המוגדרות על אותו מרחב X לפונקציית מחיר אי שלילית ($u(x)$) - נוסחה (1) במאמר. נוסחה זו נראית קצת מפחיד אבל צריך לזכור שבסקה הכל OT מוגדר עד כמה מידות הסטברות "קרובות" (כמו מרחוק [KL](#) ו-[JS](#)). המקרה הפרטי של OT שבו פונקציית מחיר הינה מרחק k_L (בין שתי נקודות x ו- y) עברו $0 < k \leq \infty$ נקרא מרחק וורשטיין מסדר k . כאשר $k=1$ הוא המרחק זהה נקרא מרחק [earth mover](#).

אך בואו נבין מה זה בעצם מרחק OT המתוואר כאמור ע"י נוסחה (1) במאמר. יש לנו משווה קצת מפחיד: מופיע שם איזה מינימום מעל כל מידות הסטברות ($u(x)$) על מרחב המכפלה (product) של X עם עצמו כאשר הפונקציות השוליות של דן הן מידות ההסטברות שעבורן אנו מחשבים את המרחק, k_L מ- P ו- Q . תחת סימן האינטגרל יש לנו את המרחק בין הנקודות. כמובן מרחק OT מוגדר כמרחק הממוצע המינימלי מעל כל ההתפליגיות דד האפשריות, המקיים את התנאי מהמשפט הקודם.

בשביל להבין את הנוסחה זו יותר טוב, בואו ניקח $k=1$ והמරחק האוקלידי כמטריקת המרחק c . בנוסף נניח שמרחב X הינו חד מימדי (\mathbb{R}). למה OT נקרא מרחק Earth Mover במקרה זהה? בעצם המרחק הזה מגדיר כמה "מסה" (הסתברותית), אנו צרכים להעביר בשביל להפוך את מידת ההסטברות P ל- Q כאשר המחר שולב נקודה x מהתומך של P לנקודה y מהתומך של Q והוא $|x-y|$. עכשו למה יש בנוסחה מינימום, אтем העברת נקודה x מהתומך של P לנקודה y מהתומך של Q הוא $|x-y|$. הדרשת כמו שאותם מבינים אפשר "להפוך P ל- Q " במספר דרכים ואנו רוצים את הדרך הכי זולה (הדורשת העברת של כמה שפחות מסה). הדבר האחרון לנו להבין בנוסחה המגדירה את OT, הוא מידת הסטברות על מרחב המכפלה של X עם עצמו? פונקציה זו מגדרה איזה "חלק" מהמסה ההסתברותית בנקודת x מהתומך של P אנו מעבירים לנקודה y מהתומך של Q . לדוגמה אם יש x ל- x הסטברות 0.5 אנו יכולים להעביר שליש ממנה של P לנו מעבירים לנקודה y מהתומך של Q . לדוגמה y_1 ו- y_2 שני שלישים $2/3 * 0.5 = 0.33$ לנקודה y_2 מהתומך של Q . התנאי שהפונקציות השוליות של דן צרכות להיות שוות ל- P - Q -OT נדרש כי אנו רוצים להעביר את כל המסה ההסתברותית מכל הנקודות מהתומך של P לכל הנקודות מהתומך של Q בלי לאבד (או להרוויח) מסה.

עזרה לגבי OT: להבדיל כמעט מכל מרחק בין מידות הסטברות, מרחק OT (וכМОון המקרה הפרטי שלו WD) לוקח בחשבון של התכונות של הסיטים עליהם מידות אלו מוגדרות בצורה מפורשת ע"י התוצאות במרקח בין הנקודות שלהם.

מציאת מרחק וורשטיין:

למרות האינטואיטיביות הרבה שיש בהגדירה של OT ו-WD בפרט, מציאת אינה טריוויאלית ברוב המקרים. עבור $1=\mathbb{1}$ ניתן להשתמש (כמו שעשו ב-GAN Wasserstein) בתצוגה הדואלית של בעית אופטימיזיה המגדירה אותה (שוויון רובינשטיין - קנטורוביץ - RK). במקום לחשב את המינימום על מידות הסתברות מעלה מרכיב המכפלה, RK מחפשת למקסם את הפרש התוחלות של $\mathbb{1}$ מעל P ומעל Q כאשר $\mathbb{1}$ היא פונקציית לפישץ עם מקדם 1.

אולם במקרה שלנו גם בעית האופטימיזיה הדואלית היא רחוקה מלהיות פשוטה לפיצוח. במקרה של שני דאטאסתים בגודל סופי ניתן להגדיר את מידות ההסתברות על המרכיבים שלהם סכום של פונקציות דלתא על הנקודות (דוגמאות) של הדאטאסתים. מרחק בין נקודות בדאטאסתים ניתן להגדיר באמצעות מטריצה כאשר איבר (j,i) שלו הוא מרחק בין נקודה j מהדאטאסט הראשון לבין i מהדאטאסט השני. ניתן לראות כי בעית אופטימיזיה (המקורית) עבור WD הופכת לעית תכנות לנארית במקורה זהה. ד"א מידת הסתברות על מרחב המכפלה $\mathbb{2}$ של עלייה מבצעים אופטימיזיה ניתנת לתיאור באמצעות מטריצה גם כן. עדין לדאטאסתים גדולים הפתרון של בעית תכנות לנארית דורש משאבי חישוב אדרירים ולא feasible. ב-2013 Sinkhorn הציע להוסף לבעה זו איבר רגולרייזציה המודד מרחק KL בין $\mathbb{2}$ לבין המכפלה הקרטזית של P ו- Q . נוספת זו מאפשרת לפטור את הבעה בזרה יותר עיליה.

מרחק בין דאטאסתים דרך מרחק ורסטיין:

נחזיר כעת לבעה שלנו ונראה איך מגדירים את מרחק בין דאטאסתים באמצעות כל המושגים שהגדכנו. קודם כל נציין כי מידת ההסתברות עבור דאטאסט מוגדרת על מרחב Z שהוא המכפלה הקרטזית של מרחב הפיצרים ומרחב הליבלים. ד"א מרחב הליבלים אינם חייבים להיות זהים עבור שני הדאטאסתים, אך נניח זאת כאן לשפטות ההסביר. המאמר מציע להגדיר את המרחק בין שתי דוגמאות: $d_{\text{OT}}(x_1, x_2) = \sqrt{\sum_i p_i(x_1) \log \frac{p_i(x_1)}{p_i(x_2)}}$

סכום של המרחקים בין x_1 ו- x_2 (השייכים לדאטאסט הראשון והשני בהתאם) ובין y_1 ו- y_2 במרחב הליבלים. בעצם המרחק מוגדר כ shores k מהракים מהמשפט הקודם.

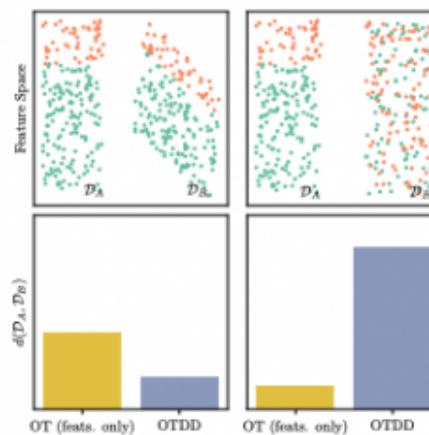


Figure 1. The importance of labels: the second pair of datasets are much closer than the first under the usual (label-agnostic) OT distance, while the opposite is true for our (label-aware) distance.

از המרחק בין הפיצרים (המרחק הראשון) מחושב בצורה ישירה (אוקלידי או כל מרחק מתאים אחר). המרחק בין הליבלים קצר יותר בעית. הדבר פשוט ביותר הוא לตาราง כל ליבל כמוצע של הפיצרים של כל הדוגמאות נשאות הליבל זהה אך זה לא מספיק מייצג את הליבל. הדרך היotta טוביה היא לחשב אונחה כמרחב ורסטיין בין התפליגיות המותנה של פיצרים בהינתן הליבלים. עם המרחק בין $\mathbb{2}$ ו- $\mathbb{2}$ מוגדר כך, ניתן להוכיח שזה מטריקת מרחק תקינה, וגם מוגדרת על סטם דיסקרטיים כמו שאנו חוצים. בסוף המרחק בין הדאטאסתים

מוגדר (בדומה ל-OT) כמינימום על כל מידות מכפלה על Z עם עצמו. את הביעיה זהו ניתן לפתור עם הוספת איבר רגולרייזציה $L1$ כמו שהזכרתי קודם. נצערנו אפילו לפתור זהה יש סיבוכיות $(n \log^5 n)$ (כאשר n הוא גודל הדאטאסט) שהופך אותו לא יישם לדאטאסטים גדולים. במקומן זאת המחברים מציעים לשערק את התפלגות המותנית של פיצרים בהינתן ליביל באמצעות גאושאננס שעבורם קיימים ביטוי סגור עבור WD . סיבוכיות החישוב במקרה הזה יורדת ל- 2^8n . המאמר גם מוכיח כי המרחק המוצע עם שערק גאושאנן זה חסום ע"י המרחק המקורי מלמעלה.

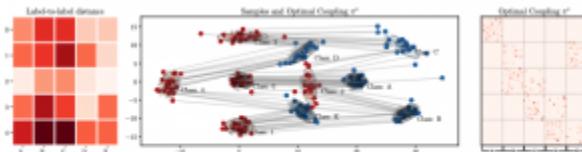


Figure 2. Our approach represents labels as distributions over features and computes Wasserstein distances between them (left). Combined with the usual metric between features, this yields a transportation cost between datasets. The optimal transport problem then characterizes the distance between them as the minimal possible cost of coupling them (optimal coupling π^* shown on the right).

הישגי מאמר:

עבור מגוון זוגות של דאטאסטים המאמר משווה את הפרשי השגיאה על טսט סט של המודל עבור הדאטאסט השמי בין שני תרחישים: אימון רגיל מאפס מול אימון של הראשן וכיוול של השמי (מאוחתל עם המשקלים של הראשן). המחברים מראים שככל שהмарחק המוצע בין דאטאסטים קטן יותר, הירידה ההפרש קטנה יותר ככלומר יותר דמיון (מרחק קטן יותר) בין דאטאסטים מתורגם ל"רמת הצלחה" בכיוול של מודל מהדאטאסט הראשן לשמי.

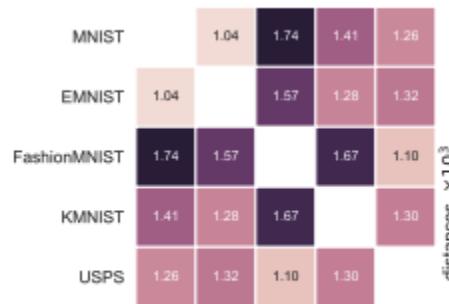


Figure 4. Pairwise OT Distances for *NIST+USPS datasets.

דאטאסטים:

MNIST, FASHION-MNIST, KMNIST, letters EMNIST

נ.ב.

מאמר עם רעיון מאד מעניין. מסקרן לראות האם גישה זו תעבור עבור דאטאסטים יותר "רציניים". נזכיר כי במרקח המתואר במאמר אין התחשבות לא בפוקציית לוס ולא בסוג המודלים שימושיים בהם לאחר מכן לסייע - לי נראה תוספת של "התחשבות" כלשהי בסוג המודלים עשויה לשפר את התוכנות של המרחק המוצע. מקווה שנראה הרחבות בקרוב.

Review 58: Gradient Descent with Early Stopping is Provably Robust to Label Noise for Overparameterized Neural Networks

פינת הסוקר:

המלצת קריאה ממיק: חובה לאלו שורצים להבין את התהליכי המתרחשים במהלך אימון של רשתות נירונים, לשאר מומלץ לעבור על המסקנות בלבד.
בahirot כתיבה: בימונית.

רמת היכרות עם כלים מתמטיים וטכניקות של DL/ML הנדרשים להבנת מאמר: הבנה عمוקה בחדו"א מתקדם ובתורת האופטימיזציה.

ישומים פרקטיים אפשריים: מאמר תיאורטי שעשוי לעזור改善 את תהליכי האימון של רשתות נירונים.

פרטי מאמר:

لينك למאמר: [זמן-can](#)

لينك לקוד: [זמן-can](#)

פורסם בתאריך: 03.07.2019, בארכיב

הציג בכנס: ICML2019

תחום מאמר:

- חקר שיטות אופטימיזציה לאימון של רשתות נירונים

כלים מתמטיים, מושגים וסימונים:

- Gradient Descent - GD
- מטריצת קווריאנס של רשת נירונים
- מטריצת קרבול של רשת נירונים

תמצית מאמר:

המאמר טוען (ומוכיח ריגורוזית) כי עצייה מוקדמת של אימון (באמצעות gradient descent) של רשתות נירונים overparameterized תורמת לרוביוטיות של הרשת המאומנת ללייבלים רועשים. המאמר בעצם מוכיח שבאייטרציות הראשונות של GD מצליח "ללמוד" איך "נראות דוגמאות שהלייבלים שלהם נכונים" ואם ממשיכים

להריז אותו, הרשת מתאימה את עצמה גם לדוגמאות בעלות הליבלים לא נכוןים. כמובן שם זה המציב, המשך אימון של רשת אחריה השלב שהוא "למדה" את הליבלים הנכונים, פוגע בביוצעה של הרשת על טסט סט (כלומר יכולת ההכללה של הרשת יורדת).

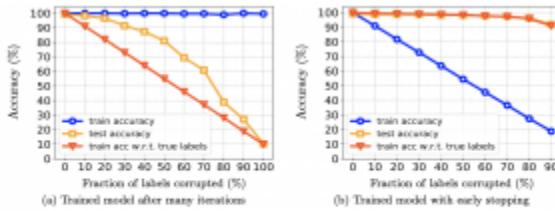


Figure 1: In these experiments we use a 4 layer neural network consisting of two convolution layers followed by fully-connected layers to train MNIST with various amounts of random corruption on the labels. In this archive the convolution layers have width 64 and 128 kernels, and the fully-connected layers have 256 and 10 out respectively. Overall, there are 4.8 million trainable parameters. We use 50k samples for training, 10k sample validation, and we test the performance on a 30k test dataset. We depict the training accuracy both w.r.t. corrupted and uncorrupted labels as well as the test accuracy. (a) Shows the performance after 200 epochs of Ada

תקציר מאמר:

למרות שמקנות של המאמר ד' ברורות וקלות להבנה הוכחתן הריגוזיות כוללות שימוש בכלים מתמטיים לא פשוטים ובהגדירות מתמטיות לא טריוויאליות. עיקב כך אטמקד בהסביר של התנאים והטענות של המשפטים האלו בסקירה זו.

נתחיל את ההסביר מהתychוסות לארכיטקטורה של הרשת המופיעה כהנחה בכל המשפטים שהוצעו במאמר.

ארכיטקטורה של הרשת ואתחול:

כל התוצאות במאמר הוכיחו לרשת דו-שכבותית (שכבה חבויה אחת) כאשר השכבה השנייה הינה קבועה ולא נלמדת (מאמנים רק את המשקלים בשכבה הראשונה). הפלט של השכבה השנייה הוא סקלר. אתחול של המשקלים הינו גאוסי (כמו ברוב המאמרים התיאורטיים בראשתות נירוניים).

הנחה על נתונים:

הנחה נוספת במאמר היא שהנקודות בדאטasset הלא רועש, המתאימים לליבלים שונים, הן מספקין רוחקות אחת מהשנייה מצד אחד ומאיידן הנקודות בעלי'אותם הליבלים (קלאלסים) מספיק קרובים (פרמטר 0_ε) לסתוריאיד של הקלאסים (בעצם הגדרה במאמר טיפה מורכבת יותר ומגדירה נקודות השיכות לכל קלואס כאחד של כמה קלוסטרים, שנקרה לו Label Set). הליבלים מוגדרים כמספרים ממשיים (!!) כאשר גם הם מספקין רוחקים אחד מהשני (פרמטר δ במאמר). דאטasset בעל תכונות אלו נקרא באופן לא מפתיע clusterable dataset. בואו נחשב מה היגיון הטמון בהגדרה זו. הרוי בחרור שכך שהסנטורואידיים (מרכזים) של הקלאלסים השונים, קרובים אחד לשני, נהיה יותר קשה לאמן רשת נירונים (או כל מסווג מסווג אחר) המבדיל ביניהם. דאטasset רועש מוגדר clusterable dataset כאשר הליבלים של אחווד נתון של נקודות מכל קלואס שונה לllibלים אקראיים.

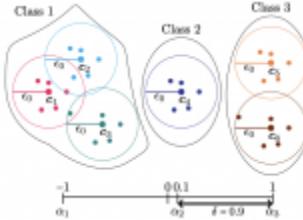


Figure 2: Visualization of the input/label samples and classes according to the clusterable model in Definition 1.1. In the depicted example there are $K = 6$ clusters, $\bar{K} = 3$ classes. In this example the number of data points is $n = 30$ with each cluster containing 5 data points. The labels associated to classes 1, 2, and 3 are $\alpha_1 = -1$, $\alpha_2 = 0.1$, and $\alpha_3 = 1$, respectively so that $\delta = 0.9$. We note that the placement of points are exaggerated for clarity. In particular, per definition the cluster center and data points all have unit Euclidean norm.

פונקציית לוס:

הרשת מאו מנת לשערר את ערך הליבל של נקודה כאשר פונקציית לוס הינה הפרש ריבועי בין פלט של הרשת לבין הליבל של הדוגמא. קרי יש לנו כאן בעיית רגרסיה עם לוס ריבועי ולא בעית סיווג.

מטריצה קווריאנס של רשת נירוניים:

זה מושג מרכז'י במאמר שבעזרתו מוכחים אם כל הטענות העיקריות. מטריצה קווריאנס של רשת נירוניים מוגדרת במאמר בתור מטריצה קרNEL אמפירית של הרשת. נזכיר כי מטריצת קרNEL של רשת נירוניים מודדת (בקירוב) את **השפעה** של צעד אחד של GD (شمудכן את המשקלי הרשת) על ערך של פונקציית הלוס של הרשת. שערוך של "מידת השפעה" זו מתבצע תוך שימוש בקירוב לינארי של פונקציית לוס שהופך למדוק יותר ככל שגדלי השכבות של הרשת גדלות יותר.

כאמור אם יש לנו שני קלסיטרים של נקודות (ולא רועשים) בעלי ליבלים שונים שקרובים אחד לשני אך רשת "צריכה לעבוד קשה" בשביל להבחן ביניהם (או במלים אחרות "לבנות" משטח המפריד בין הקלסיטרים). אז מטריצת קווריאנס C באה לעזרך לנו לכתם את היכולת הזו (הבחנה בין נקודות בעלות ליבלים שונים) של רשת נירוניים נתונה ווט של מרכזי קלאסיטרים (סנטרואדים) נתון עבור כל ליבל. המאמר מראה כי ניתן לעשות זאת באמצעות condition number condition (יסומן בהמשך ב- cond) של C. אזכיר כי cond של מטריצה מוגדר כיחס בין ערך העצמי הגדל ביותר לבין הערך העצמי הקטן ביותר של המטריצה. ככל ש-cond של מטריצת קווריאנס של הרשת נמוך יותר, אז קל יותר לרשת להבחן בין קלאסיטרים שונים.

פינת האינטואיציה: נניח כי יש שני מרכזיים של קלסיטרים של דוגמאות, הנושאים ליבלים שונים, נמצאים באותה נקודה. קל לראות שבמקרה זהה למטריצת קווריאנס יהיו שורות תלויות כלומר יהיה לה ע"ע 0. لكن cond שלה יהיה אינסוף שמסתדר עם טענה המנוסחת לעלה.

טענה עיקרית 1 של המאמר: בהינתן דאטוסט עם אחוז ליבלים רועשים נמוך מספיק, וגודל השכבה הנלמדת מספיק גדול, קרי $K * 4^{\text{cond}(2)} O$, קיימים קצב למידה (שהוא גם תלוי ב-cond של מטריצת קווריאנס) שעבורו, אחרי מספר צעדי GD, הרשת תלמוד לזהות נכון את הליבלים של כל הדוגמאות בדאטוסט. K מסמן את מספר הקלאסיטרים ב-set label (אזכיר של set label מורכב מכמה קלסיטרים של דוגמאות). מספר צעדי GD עד ההגעה לזהוי מלא של כל הנקודות הלא רועשות הוא (K) O . בנוסף המרחק המקיים בין משקלי האתחול של רשת לבין המשקלים בכל האיטרציות של אימון (עד ההגעה למצב שהרשת מזיהה נכון את כל הדוגמאות עם הליבלים הלא רועשים) יהיה נמוך יחסית לעומת הרשת "תטיל" בסביבה די קטנה סביב משקלי האתחול במהלך האימון כדי לזהות נכון את הליבלים המקיימים.

טענה עיקרית 2: עכשו נשאלת השאלה מה קורה אם אנחנו לא עוצרים את האימון מוקדם וממשיכים לאמן את הרשת עם GD. המשפט העיקרי השני במאמר נותן מענה לשאלת זו. המשפט זהה מוכיח שתחת אותם התנאים

על ארכיטקטורת הרשת ועל מבנה של דאטאסט, בשביל לתת דיק של 100% על דאטאסט עם לייבל רושע אחד (לפחות נכון כל הדוגמאות כולל זו נשאת הליבל הרושע) המרחק שהמשקלים של הרשת צריכים לעבור (קרי המרחק בין המשקלים ההתחלתיים לבין אלו של הרשת המאומנת) צריך להיות לפחות $\epsilon/3$, כאשר המוניה מוגה חסם על המרחק בין ערכי הליבלים השונים, והמננה מתאר את הרדיוס המקסימלי של הקולסטררים של אותם הליבלים. ככל שהקולסטררים של דוגמאות יותר גודלים (המננה עולה) אז המרחק מתקצר (הקולסטררים יותר מרוחים וקל למצאו וקטור משקלים מסווג נכון הליבל הרושע). כאשר המרחק בין ערכי הליבלים עולה (המננה עולה) המרחק שמשקלים צריכים לעבור מתרוך ("מכיריהם את הרשת לטעות גם כישש לה ביטחון גובה"). הכל תחת אתחול גאוס של המשקלים.

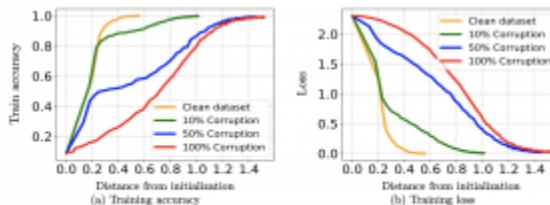


Figure 3: We depict the training accuracy of a LENET model trained on 3000 samples from MNIST as a function of relative distance from initialization. Here, the x-axis keeps track of the distance between the current and initial weights of all layers combined.

כליים מתמטיים המשמשים להוכחות:

המאמר משתמש ביקובין L של מיפוי המודל באמצעות הרשת (עבור דאטאסט נתון) בשביל לנתח את מטריצת קוווריанс של הרשת. המחברים הציגו את השארית הרועשת (הפרש בין פלט של רשת לבין לייבל רושע) סכום של השארית הנקייה והרועל באופןו לייבל רושע. לאחר מכן הם הוכיחו שהשארית הנקייה "מכוסה" ע"י תת-מרחב של מרחב העמודות של L המתאים לערכים סינגולריים גדולים של L. זה למעשה מאפשר לרשת ללמוד את הליבלים הנקיים במחירות (גרדיינטים חזקים). לעומת זאת "הרועל בליבל" עצמו מכוסה ע"י תת-מרחב המתאים לערכים סינגולריים קטנים ששופקsha על האימון של הליבלים הרועשים (גרדיינטים חלשים). לדעתם זו מסקנה מאוד חזקה.

דאטאסטים:

MNIST, CIFAR10

.ג.ב.

מאמר מאד חשוב העוזר להבין את האופן שבו רשותות "לומדות" את הדאטה.

Review 59: Unsupervised Discovery of Interpretable Directions in the GAN Latent Space

פינט הסוקר:

המלצת קריאה ממייק: מומלץ לעוסקים ב-GANs לשאר רק אם יש זמן פנוי.

bahiorot כתיבה: גובהה.

רמת היכרות עם כלים מתמטיים וטכניקות של DL/ML הנדרשים להבנת מאמר: היכרות עם עקרונות של GANs מספקת.

ישומים פרקטיים אפשריים: מציאת כיוונים במרחב הלטנטי הגורמים לשינוי של מאפיין ויזואלי בודד של התמונה המוגנרטת.

פרטי מאמר:

[לינק למאמר: זמן כאן](#)

[לינק לקוד: זמן כאן](#)

פורסם בתאריך: 24.06.2020, בארכיב

הוזג בכנס: ICML 2020

תחום מאמר:

- GANs
- חקר של המרחב הלטנטי של GANs

כלים מתמטיים, מושגים וסימונים:

- וקטור (כיוון) בר פירוש (interpretable direction).

תמצית מאמר:

המאמר הנזכר מציע שיטה למציאה של וקטורי (כיוונים) "בר פירוש" (interpretable directions) במרחב לטנטי של GAN מאומן. וקטור בר פירוש int_v מוגדר ככזה שהוספטו לכל וקטור v מהמרחב הלטנטי של GAN מאומן, שהשינו בין התמונות המוגנרטות באמצעות וקטורים אלה ($v - \text{int_v} + v$), יהיה במאפיין ויזואלי אחד בלבד של כגון צבע גוון עור, צורת גבורה, רקע וכדומה. השיטה המוצעת לא תליה בארכיטקטורה של GAN ולא דורשת שום *supervision* (!!).

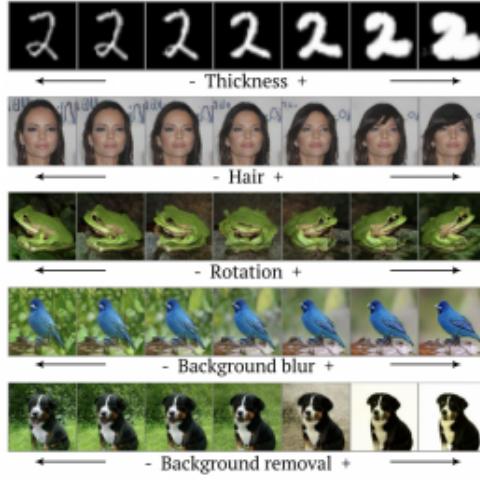


Figure 1. Examples of interpretable directions discovered by our unsupervised method for several datasets and generators.

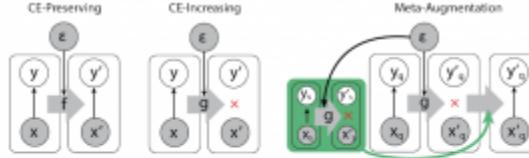


Figure 2: **Meta-augmentation:** We introduce the notion of *CE-preserving* and *CE-increasing* augmentations to explain why meta augmentation differs from standard data augmentation. Given random variables X, Y and an external source of random bits e , we augment with a mapping $f(e, X, Y) = (X', Y')$. **Left:** An augmentation is *CE-preserving* if it preserves conditional entropy between x, y . **Center:** A *CE-increasing* augmentation increases $H(Y'|X')$. **Right:** Invertible CE-increasing augmentations can be used to combat memorization overfitting: the model must rely on the base learner to implicitly recover e from x'_g, y'_g in order to restore predictiveness between the input and label.

רעיון בסיסי:

הנחת יסוד של המאמר אומרת שכיוונים ברוי פירוש שונים גורמים לטרנספורמציות בעלות שני רבי של התמונה, ככלומר יכולים שניין להבחין בין טרנספורמציה אחת לאחרת בקלות. לכן בתהילר הלמידה המחברים מנסים לאתר (ללמידה) כיוונים במרחב הlatentי הגורמים לטרנספורמציות שונות מאוד בתמונות הנוצרות.

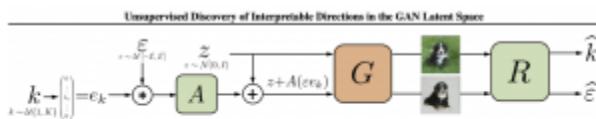


Figure 3: Scheme of our learning protocol, which discovers interpretable directions in the latent space of a pretrained generator G . A training sample in our protocol consists of two latent codes, where one is a shifted version of another. Possible shift directions form a matrix A . Two codes are passed through G and an obtained pair of images go to a reconstructor R that aims to reconstruct a direction index \hat{k} and a signed shift magnitude \hat{v}_k .

תקציר מאמר:

כפי שצינו במאמר מחקרים המתמקדים בחיפוש כיוונים ברוי פירוש במרחב latentי של GAN מערבים supervision כלשהו בתהילר החיפוש. מה שנהוג לעשות הוא לבצע טרנספורמציות ברוח פירוש לתמונות (סיבוב, הקטנה, הוספה משקפים וכדומה) ולראות איזה כיוונים במרחב latentי גורמים לשינויים האלה. קו נוסף של מחקר

בנושא זה מתמקד במבנה של גאנים בעלי פיצרים לא מעורבים (disentangled) שהם ניתנים להפיכת כיוונים בר' פירוש יחסית בקלות. המאמר מצין אימון גאן עם פיצרים לא מעורבב זו משימה קשה (שזה נכון) וגם טווען שההטוצאות של מודלים כאלה לא מרשים במיוחד יחסית ל-SOTA. InfoGAN, OoGAN, StyleGAN-2 ודוגמאות של ID-GAN�� אובל משום מה לא מתייחס למושג זה. סגנון חדשם לפני פרסום המאמר הנסקר. דרך נוספת לנתח כיווני בר' פירוש בגאנים היא לחזור ו"ע של מטריצת יקובין של הGENERATOR.

המאמר מציע שיטה "ישירה" יותר לפתרון של בעיית חיפוש כיוונים בר' פירוש במרחב לטנטי של GAN. השיטה המוצעת לוקחת רשות GENERATOR מאומנת ומנסה למצוא K (שווה בדרך כלל זה מינימום של ה-LOSS של הGENERATOR) וקטורים בר' פירוש במרחב הלטנטי. החיפוש נעשה בדרך מאד פשוטה האינטואיטיבית. מאומנים רשות כאשר וקטורים בר' פירוש הם חלק מהמשקלים שלהם. למעשה המחברים לא עושים שום טרנספורמציה מפורשת לתמונה כמו שנוהג לעשות במאמרים הקודמים אלא רק "משחקים" עם המרחב הלטנטי. בגדול המאמר מציע להציג וקטור z מהמרחב הלטנטי של הגאן המאמן ולקטור לוגרורי בר' פירוש מאומן v . לאחר מכן מגנרטים שתי תמונות באמצעות הזנות $z - v + z$ לגנרטור מאומן. בשלב האחרון מזינים את התמונות הללו לרשות שמנסה לשערך מהו הכיוון של וקטור v המוסף $-z$.

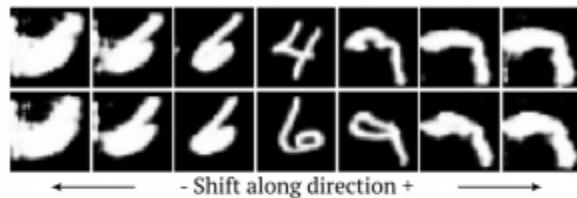


Figure 4. Direction of collapsing variation. The regression term in our objective prevents discovery of such directions.

כעת ניתן תיאור יותר מפורט של השיטה המוצעת:

אימון הרשות למציאת וקטורי בר' פירוש:

1. מגרילים את מספר ה- c_i בין 1 ל- K ויוצרים מזה וקטור hot_i .
2. מגרילים גודל (אורך) a של וקטור זה (מהתפלגות יוניפורמיית סימטרית סביב 0).
3. מכפילים הוקטור $ah = v$ במטריצה A עם משקלים מאומנים (A למעשה מכילה וקטורי בר' פירוש שאנו חסם מנסים לשערך/ללמוד (בדומה לבנייה של וקטורי יציג למילון נתון ב-word2vec)).
4. מגרילים וקטור z מהמרחב הלטנטי שכך גאנ מהתפלגות גאומטרית סטנדרטית.
5. מגנרטים שתי תמונות, הראשונה $m-z$ והשנייה $m-v+z$ באמצעות הזנות לגורטור המאמן.
6. מעבירים שתי תמונות אלו דרך רשות עם משקלים מאומנים R כאשר מטרתה לשערך את מספר ה- c_i ואת הגודל שלו a (ז"א הפלט של R הוא וקטורי הסתברויות K -dimensionי ומספר ממשי).

חשוב להבין שהמשקלים של R והמשקלים של מטריצת ה- c_i מאומנים בלבד. עקב לכך כדי תחיליר האימון העמודות של מטריצה A מנוסות "לפשט" את בעיית הסיווג שהרשות R מנסה לפתור, "באמצעות התכונות" לכיוונים קלים יותר להבחנה.

פונקציית loss: פונקציית loss מורכבת משני מחוברים: הראשון הוא איבר קרוס-אנטרכופי סנדראטי על השערור של מספר ה- c_i והשני הוא הלווי הריבועי על השערור של a . האיבר השני מהו רגולרייזציה המיועדת לכפות על אורך של וקטורי ה- c_i להשפיע באופן רציף על התמונה המוגנרטת במטרה למינע מיפוי של כל ה- c_i לקבוצה קטנה של תמונות.

הערה לגבי מטריצת הcyonim: המחברים ניסו לבחור מטריצת cyonim משתי צורות - בעלת עמודות עם אורך 1 ומטריצה אורתונורמלית (עמודות אורתוגונליות בעלות אורך 1). עברו דאטסהטים שונים צורות שונות של מטריצת cyonim הציגו ביצועים יותר טובים מהשתפים שנבדקו, אך לא מצאתי התיחסות או דיון בסוגיה זו במאמר.



Figure 7. Examples of directions discovered for Spectral Norm GAN and AnimeFaces dataset.

הישגי מאמר:

בסיומו של דבר חלך הcyonim (לא מצאתי מה האחוז) שהתקבלו כתוצאה מהתהליך הזה אכן נמצאו כגורם לשינויים במאפיין אחד של התמונה. שינויים אלו ניתנים להבhana ע"י עין אנושית (כגון צבע שער, סיבוב של תמונה, גוון עור, אודם וכדומה). דבר מעניין שהמחברים הצליחו לגלות הוא שאחד הcyonim שהם מוצאים הוא אחראי על הרקע של התמונה שאיפשר להם לטען שהם מצאו דרך לעשות אוגמנטציה טובה לדאטסהטים למגוון שימושות. מעניין שהמאפיינים הייזואליים שמתאימים לכיוונים בררי פירוש שנמצאו, משתנים (!!?) בין מודלי GAN לבין דאטסהטים שונים. אצין כי בנוסף להבhana האנושית, מטריקה נוספת נוספה לשערוך ביצועים שהמאמר השתמש בה היא דיק שחרור (RCA) הcyon k בראשת R. כמון ש-RCA גבוהה לא מUID על כך שמצוינו cyon בר פירוש חזק כי קיימים שילובים של כמה מאפיינים בתמונה קלים להבhana.

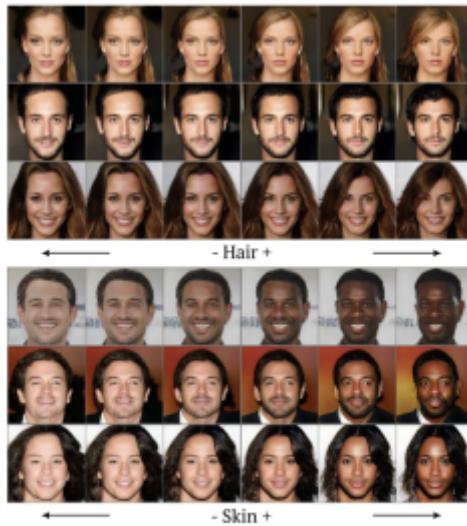


Figure 8. Examples of directions discovered for ProGAN and CelebA dataset.

דאטסהטים: MNIST, AnimeFaces, Imagenet and CelebA-HQ

סוגי GAN שנבדק: Spectral Norm GAN, ProGAN, BigGAN

ג.ב. אמר עם רענן נחמד, אך קצת לא מבושל. מעבר לאינטואיציה הבסיסית לא מצאת הסבירים למה הרעיון שלהם עבד. גם הייתי רוצה לראות דיוון עמוק יותר בתלות בין הכוונתי ברי פירוש שנמצא בין הדאטסהט שעלי אומן GAN ובארכיטקטורה של GAN. בקיצור נראה שהכוון הזה רק בתחילת דרכו ומקווה לראות את המשך.

Review 60: Diffusion Models Beat GANs on Image Synthesis

פינת הסוקר:

המלצת קריאה ממייק: חובה למי שרצה למדוד מודלים גנרטיביים פרט לגאנים ול-VAE.

בahirot cttiba: ביןונית.

רמת היכרות עם כלים מתמטיים וטכנולוגיות של DL/ML הנדרשים להבנת מאמר: הבנה טובה של עקרונות VAE, הבנה של שיטות דוגימה מתקדמות כמו [динמיקה של לנגן](#).

ישומיים פרקטיים אפשריים: ייצור תמונות יותר "איכותיות" מהగישות המתחזרות, קרי גאנים ו-VAE.

פרטי מאמר:

[לינק למאמר: זמין להורדה.](#)

[לינק לקוד: זמין כאן](#)

פורסם בתאריך: 21.06.21, בארכ'יב.

הציג בכנסו: טרם ידוע.

תחומי מאמר:

- מודלים דיפוזיוניים כЛОMR Diffusion Denoising Probabilistic Models - DDPM לגנרטוט של>Data ויזואלי.

ידע מוקדם:

- הבנה טוביה בטכניקות מבוססות variational inference לניתוח פונקציות נראות מירבית (כמו ב-VAE).
- רקע טוב בהסתברות לא-Յיזיק ()

מבוא:

מודלים גנרטיביים מבוססי רשותות נוירונים לייצרת>Data ויזואלי רשמו התקדמות מרשימה בשנים האחרונות. מודלים כמו [StyleGAN2](#) ו- [Q-VAE2](#) מסוגלים לגנרט תמונות מגוונות באיכות מרשימה בדומיננסים שונים. וכך גם רוב המודלים הגנרטיביים עם תוצאות SOTA הם מסוג גאן ו-VAE (עם יתרון ניכר לגאנים). מלבד גאנים ו-VAEs קיימים סוגים נוספים של מודלים גנרטיביים מבוססים על גישות אחרות כמו מודלים דיפוזיאוניים ומודלים מבוססי זרימה (flow). עד כה מודלים אלו לא הצליחו (לפחות מבחינת המדדים המקובלים כמו FID ו-Inception Score) להציג ביצועים ברι השוואה עם תוצאות SOTA. נציין כי לפחות מבחינה ויזואלית איות התמונות הנוצרות באמצעות מודלים דיפוזיאוניים ומובוסטי זרימה לא נופלת מזו של אלו הנוצרות באמצעות גאנים ו-VAE-ים המתקדמים ביותר (דעה אישית).

המאמר הנסקר הוא הראשון (למייט ידיעתי) שבו הצליח מודל דיפוזיאני להגיע לביצועים טובים יותר ממודלים גנרטיביים, אשר נתונים כiom את התוצאות הטובות ביותר. זו בשורה משמעותית עד כדי כך שמחברי המאמר ציינו אותה *ישירות בכותרת*:



Figure 1: Selected samples from our best ImageNet 512×512 model (FID 3.85)

תמצית מאמר:

המאמר הנסקר מתבסס על שני מאמרים קודמים ומציג שורת שיפורים שהצליחו "להרים את הביצועים של DDPM" לרמה של גאנום ומעבר לכך:

- [מאמר רקע 1](#) למעשה הציע את מה שנקרא Denoising Probabilistic Model או בקיצור DDPM. מעניין כי מודלים דיפוזיאוניים לגנרטוֹן דатаה הומצאו עוד ב-2015 ב- [מאמר רקע 0](#).
- [מאמר רקע 2](#) הציע רפורטורייזציה של פונקציית הלוס, שינוי של תהליכי האימון (יפורט בהמשך) וכמה טרייקים נחמדים נוספים שבפועל שיפורו את יכולות התאמנות המגוננות באמצעות המודל.
- המאמר הנסקר מציע דרך לנצל דатаה מותוג לאימון מודל דיפוזיאוני לצד כמה שיפורים ארכיטקטורתיים של רשתות ניירונים המעורבות בתהליכי הغانרטוֹן.



Figure 6: Samples from BigGAN-deep with truncation 1.0 (FID 6.95, left) vs samples from our diffusion model with guidance (FID 4.59, middle) and samples from the training set (right).

כאמור, המאמר הנסקר מציג שורת שיפורים [למאמר רקע 2](#) שבעצמו מהווה גרסה משודרגת של [מאמר רקע 1](#). עקב לכך אתחיל מסקירה מפורטת ועמינית של מודל דיפוזיאוני שהוצע [במאמר רקע 1](#), לאחר מכן אסקור את השדרוגים של [מאמר רקע 2](#) של המאמר הנסקר.

תקציר מאמר רקע 1:

מודל דיפוזיאוני DDPM לגינרוט DATAה: הרעיון של DDPM הוא די פשוט. לוקחים תמונה, מוסיפים אליה רעש גausi במשר כמה איטרציות (מאות או אלפיים) עד שהתמונה הופכת להיות לרעש גausi איזוטרופי ($I(0, N)$ - זה נקרא תהליך קדמי (forward process). המטרה של מודל דיפוזיאוני הוא למדל (ללמוד) את התהליך ההופיע (reverse process) - כלומר לגנרט תמונה מרוש גausi איזוטרופי צעד אחריו צעד.

מטרת אימון DDPM: המטרה היא למדל את התפלגות $(x_t | x_{t-1})$ כאשר x היא התמונה המתקבלת באיטרציה t של התהליך הקדמי המתואר לעיל. באופן פורמלי, אם נסמן את התפלגות התמונות מהדאטסט ב- $q(x_0)$, אז התהליך הקדמי יתואר באופן הבא:

$$q(x_1, \dots, x_T | x_0) := \prod_{t=1}^T q(x_t | x_{t-1}) \quad (1)$$

$$q(x_t | x_{t-1}) := \mathcal{N}(x_t; \sqrt{1 - \beta_t} x_{t-1}, \beta_t \mathbf{I}) \quad (2)$$

כאשר $(1, 0) \in \beta$ הם סדרה של קבועים דטרמיניסטיים. T מסמן את מספר האיטרציות של התהליך.

חיזוי: כמו שכבר נזכר ניחשתם זמן החיזוי הוא עקב האכילים של מודלים דיפוזיאוניים. כדי לבנות תמונה מרוש אנו צריכים לשחזר את כל הצעדים של התהליך ההופיע. המאמר הנסקר מדבר על בערך 4000 איטרציות המצריכים הריצה של 4K רשותות אחת אחרי השניה שזה כפונן מאוד בעיתתי.

פרטים על הרעש המוסף: תוחלת (פר פיקסל) של רעש גausi המוסף בכל איטרציה תלויות בערך של הפיקסל. רעש המוסף עבר פיקסל $\{j\}$ באיטרציה t מוגדר באמצעות התפלגות נורמלית $\mathcal{N}(\alpha_t, \beta_t)$, כאשר $\beta = 1 - \alpha$. x_t הינו ערך הפיקסל $\{j\}$ בתמונה מושפעת מאייטרציה $t-1$.

נקודה חשובה: מידול של התהליך ההופיע עשוי להיראות פשוט לאור העובדה שההתהליך הקדמי (המתואר באמצעות התפלגות $(x_t | x_{t-1})$) מתפלג גausiot. אולם השערוך של $(x_t | x_{t-1})$ **אינו ממשימה פשוטה** והתפלגות זאת אינה גausiot. הסיבה לכך היא שלהבדיל מההתהליך הקדמי שהוא הוספה של רעש גausi בעל תוחלת ושונות ידועות לתמונה, התהליך ההופיע הוא למעשה ניקוי של תמונה מושפעת מחלוקת של הרעש שיש בה (מכאן באה המילה denoising בשם של המודל). כדי לבצע צזה נדרשות "הבנות" של התפלגותות של תמונות המתקבלות בשלבים השונים של תהליך דיפוזיאוני.

יעקב המורכבות התמונה במידול של $(x_t | x_{t-1})$, משערכים אותה באמצעות התפלגות גausiot פרמטרית $\mathcal{N}(x_t | \theta)$ מהמודל ע"י, מי היה מנחש, רשת נירונית. פורמלית:

$$p_\theta(x_{t-1} | x_t) := \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \quad (3)$$

כך $I_t \gamma = (x_{t-1} | x_t)^\theta \Sigma$ (כלומר הרשות חוזה רק את סקלר γ).

נקודה חשובה: למה ניתן לקרב $(x_{t-1} | x_t)^\theta q$ באמצעות $(x_{t-1} | x_t)^\theta p$ **גאוסי** בדיק טוב?

הרי כבר אמרנו ש- q "טומנת בה ידע על התפלגות התמונות" של הדאטסהט עליו מאומן DDPM. מתרבר כי קירוב זה עובד טוב כאשר הרעש המוסף בכל שלב של תהליך קדמי הוא בעל תוחלות ושוויות נוכחות מסוים (אחד מהמאמר רקע מצין כי קיימת הוכחה של גאוסיות תחת תנאים מסוימים על התפלגות של $(x_t)^\theta q$ אך לא ראייתו אותה).

DDPM מול מודלים גנרטיביים אחרים: ברמת העיקרון DDPM דומה למודלים גנרטיביים אחרים כמו גאן VAE או מודלי זרימה שגם יוצרים תמונה מרושע. אבל כאן הדמיון בין גישות אלו נגמר כי הדריכים בהן הן מדלות מיפוי מרושע לתמונה הן מאוד שונות (למרות ש-VAE ו-DDPM משתמשים ב-ELBO לבניה של פונקציית המטרה שלהם).

איך מאמנים מודל דיפוזיוני? מטרת האימון של מודל דיפוזיוני היא מיקסום לוג של נראות מירבית (\log likelihood) של הדאטסהט ביחס לוקטור פרמטרים θ . כמובן לוג של נראות מירבית של דאטסהט נתון הוא סכום של $(x_t)^\theta q$ עבור כל התמונות x מהדאטסהט. בדומה ל-VAE (אך עם קצת סיכון עקב איטרציות רבות המעורבות בתהליך), משתמשים בהתאם (ELBO) כדי לקבל את פונקציית מטרה L של בעית אופטימיזיה עבור מודל דיפוזיוני:

$$\begin{aligned} L_{\text{vlb}} &:= L_0 + L_1 + \dots + L_{T-1} + L_T \\ L_0 &:= -\log p_\theta(x_0 | x_1) \\ L_{t-1} &:= D_{KL}(q(x_{t-1} | x_t, x_0) || p_\theta(x_{t-1} | x_t)) \\ L_T &:= D_{KL}(q(x_T | x_0) || p(x_T)) \end{aligned}$$

כך $(x_t)^\theta q$ הוא רעש גאוסי איזוטרופי.

הסבר על האיברים של L_{vlb} :

- L_1 - מודד עד כמה "סביר" לקבל את התמונה המקורית x_0 מתמונה x_1 שהתקבלה בשלב לפני האחרון של התהליך ההפוך.
- $L_{t-1} < T < 0$ - מודד דמיון בין ההתפלגות המשוערת $(x_{t-1} | x_t)^\theta q$ לבין "התפלגות האמיתית" $(x_{t-1} | x_0)^\theta q$ הנדגמת לתמונה x_0 מהדאטסהט.
- L_T - מודד עד כמה x , המתקיים בשלב האחרון של התהליך הקדמי, "קרובה" (במשמעותו התפלגות) לרעש גאוסי איזוטרופי.

תהליך אימון של DDPM בגודל: פונקציית הלווי שלנו היא סכום של T מחוברים אי שליליים. כדי למנוע אותה, דוגמים $T \leq t \leq 0$ ומבצעים איטרציה של gradient descent על האיבר θ -t של הסכום. כאמור אנו מאמנים רשות N כדי לחזות את התפלגות $(x_{t-1})_t$ μ לכל $t < T < 0$. בכלל איטרציה מאפטמים את הפרמטרים של N_θ כדי למנוע את הלווי L_t עבור t הנדגם (t מוזן לתוך הרשות).

פלט של הרשות: הדרך הטבעית היא לאמן את הרשות לחזות את $(x_t)_{\theta} = I_t$ γ Σ תוחלת ומטריצת הקוריאנס של $(x_{t-1})_t$ μ . אך ניתן גם לאמן N לחזות עוד פרמטרים המעורבים בתהליך (כמו התפלגות $(x_0)_t$ של התמונה המקורית x) מהם (יחד עם x) ניתן לגזר את $(x_t)_{\theta}$ μ ו- γ .

הערה: γ לא נחזה באמצעות רשות ניירונים במאמר רקע 1 אלא משתמשים בקירוב שלו - הסיבות לכך יפורטו בהמשך.

מאמר רקע 1 בחר לאמן N כדי לחזות **פרמטר אחר** אחר שנייתן לגזר ממנו את $(x_t)_{\theta}$ μ תוך שימוש בתכונות של התהליך הקדמי. כתע נרחב איך ניתן לעשות זאת. ניתן לבצע את רפרמטריזציה הבאה להתפלגות μ :

$$q(x_t | x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t} x_0, (1 - \bar{\alpha}_t) \mathbf{I}) \quad (8)$$

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon \quad (9)$$

כאשר $\bar{\alpha}_t = \prod_{j=0}^t \alpha_j$ והוא רעש גauss סטנדרטי $(0, 1)$. אינטואיטיבית די ברור כי x_t מתפלג גaussית כי x_0 נבנה מ- x באמצעות הוספת רעשים גaussים בעלי תוחלות ושינויות ידועות. בנוסף מתקיים:

$$\tilde{\beta}_t := \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t \quad (10)$$

$$\tilde{\mu}_t(x_t, x_0) := \frac{\sqrt{\bar{\alpha}_{t-1}} \beta_t}{1 - \bar{\alpha}_t} x_0 + \frac{\sqrt{\bar{\alpha}_t} (1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} x_t \quad (11)$$

$$q(x_{t-1} | x_t, x_0) = \mathcal{N}(x_{t-1}; \tilde{\mu}(x_t, x_0), \tilde{\beta}_t \mathbf{I}) \quad (12)$$

מאמר רקע 1 מאמנים N לחזות את רעש ϵ המוסף בשלב t (המחברים טוונים שזה משפר את איכות התמונות המיוצרות) שמננו ניתן לגזר $(x_t)_{\theta}$ μ באופן הבא:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right) \quad (13)$$

למעשה פונקציית לוי שהרשות N_θ מאומנת למנוע היא:

$$L_{\text{simple}} = E_{t,x_0,\epsilon} [||\epsilon - \epsilon_\theta(x_t, t)||^2] \quad (14)$$

הערה: כמו שכבר ציינתי המאמר לא משער γ אלא משתמש רק בקירובו $\bar{\beta}_t$ (שונות של $x_{t-1}|x_t$). למעשה כי $\bar{\beta}_t < \gamma$ (ערכים דטרמיניסטיים) אך בפועל שימוש בכל אחד חסמים אלו הוביל לתוצאות מאוד דומות. צריך לציין שימוש ב- L_{simple} שקול למשך המחברים בפונקציית המטרה המקורית L_{vlb} (זה נבע מהצורה של מרחק KL בין התפלגויות גאוסיות).

ארכיטקטורת רשות:

mbosst על זו של [PixelCNN++](#) שהוא שילוב [U-Net](#) ו- [Wide ResNet](#). כדי לקודד את מספר איטרציה t משתמשים בקידוד מיקומי (positional encoding) מהמאמר המקורי על הטרנספורמרים (All is Need You, זכריהם?). המחברים גם משתמשים במנגנון self-attention בין שכבות קונבולוציה ברזולוציות שונות.

בכך סיימנו לתאר את DDPM כמו שהוזג [במאמר רקע 1](#). עת נעבור לשינויים שהוצעו למודל זה [במאמר רקע 2](#) ובמאמר הנסקר.

תקציר שיפורים/שינויים של DDPM:

למעשה יש ארבעה סוגים של שיפורים שבזכותם DDPM הצליח להחות את הגאנים:

שינויים בפרמטרים של התהילה המקורי:

- [מאמר רקע 2](#) (פרק 3.2): קבועי $T \leq t \leq 0$, $\bar{\beta}_t$ קבועים באופן שונה. המחברים שמו לב כי השלבים האחרונים של התהילה המקורי יוצרים תומנות רועשות מדי ולא תורמים לאיכות התמונה המוגנרטת. עקב כך הוצע לקבוע קבועים אלו כדי "להאט הפיכתה של תמונה לרעש".

שינויים בפונקציית לוס ובתהילה אימון של N_θ :

- [מאמר רקע 2](#) (פרק 3.1): כאמור בגרסתה המקורי של DDPM המחברים החליטו לא לשער שונות γ_t של $x_{t-1}|x_t$ והסתפקו בשערוך של תוחלתו (באופן עקייף דרך ϵ). ההסבר שלהם לגבי למה זה עובד מספיק טוב היה טמון בעובדה כי $\bar{\beta}_t < \gamma_t < \bar{\beta}_t$ אך $\bar{\beta}_t$ הם מאוד קרובים עבור רוב ערכי ϵ . [מאמר רקע 2](#) נקט בגישה אחרת והציג רפרמטריזציה קמורה של $\log \bar{\beta}_t = \exp(v \log(\bar{\beta}_t) + (1 - v))$ כאשר $v \in (0, 1)$ הוא אימנו רשות לשערוך של ϵ . נציין כי פונקציית הלוס הקודמת L_{simple} לא מכילה את γ_t אך המחברים השתמשו צירוף לינארי של L_{simple} ו- L_{vlb} בתור פונקציית לוס חדשה.
- [מאמר רקע 2](#) (פרק 3.3) מבהיר דוגמה יוניפורמיות ב- ϵ -based importance sampling. ההסתברות של בחירת ערך ϵ פרופורציונלית לערך $\bar{\beta}_t$ המוצע. לטענת המחברים זה מקטין את התונודתיות של הגראדיינטיהם שלהם.

- המאמר הנסקר משתמש בדатаה מתויג לאימון של DDPM. הרעיון הוא לנצל תמונות מתויגות ל”ניוטוט של מודל דיפוזיוני לכיוון” שבו תמונות שהוא מייצר בתהיליך הופci, יסוגו עם הקטגוריה נוכנה בוודאות גבוהה באמצעות מסוג אמיתי. ככלומר לכל ערך של t מאמנים רשות מסוגת $N_{\phi,t}$ שהפלט שלו עברו תמונה x_t (מייטרציה t) הוא $(x|y)_\phi$ בעבר קטגוריה y . במהלך האימון לתמונה בעלי קטגוריה y , “מתקנים” את התפלגות $x_{t-1}|x_t$ באופן כזה שההתמונות תקבלנה ערך גובהה של $(x|y)_\phi$. במקום לשערך $(x|y)_\theta$ k אנו משערכים (דוגמים מ-): $(x|y)_\phi k(x|y)_{t-1} = Z p_\theta(x|y)_{t-1} = (y|x_{t-1})_\theta$. כמו שאתם יכולים לנחש שערוך כזה לא לגמרי כל ומערב מתמטיקה לא טריואלית (זה מבוסס על הקשורים [לдинמיקה של לנגבין](#)). יותר פרטים נמצאים בפרק 4 של המאמר הנסקר.



Figure 3: Samples from an unconditional diffusion model with classifier guidance to condition on the class "Pembroke Welsh corgi". Using classifier scale 1.0 (left; FID: 33.0) does not produce convincing samples in this class, whereas classifier scale 10.0 (right; FID: 12.0) produces much more class-consistent images.

шиפורים באררכיטקטורה של N_θ : (המעוניינים בינהם)

- מנגן **attention** בעל רזולוציות מרובות (multi-resolution).
- שימוש בבלוקים residual של BigGAN - up/downsampling.
- Adaptive group normalization (AdaGN)

זריז תהיליך החיזוי: שינוי בהגדרת תהיליך הופci שמאפשר חיזוי מדויק של $x_{t-1+m}|x_t$ עבור $0 < m < t-1$. שינוי זה מאפשר לדגם את x_t כל m צעדים ול- m גדול מקטין את זמן החיזוי באופן שימושותי. המתמטיקה העומדת מאחורי ההגדרה החדשה זו די לא טריואלית ובנוסף הטהיליך הקדמי מאבד את המרקוביות שלו כי x_t תלוי באופן מפורש גם ב- x_0 ועוד ב- x_{t-1} .

הישגי המאמר:

כאמור המודל הדיפוזיאוני המוצע הצליח להוכיח את הגאנים המוביילים מבחינת FID. זמן החיזוי עדין נותר ד' גביה יחסית לagan אבל יש שיפור ניכר יחסית למודלים דיפוזיאוניים קודמים.

ג.ב.

מאמר ממש מגניב המציר הבנה עמוקה של 3 מאמרים שקדמו לו בנושא של מודלים דיפוזיאוניים (ועוד שניים בנושאים סמוכים). המתמטיקה לא טריואלית אבל היה שווה את המאמץ.