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

---

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

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

---

## פינת הסוקר:

**המלצת קריאה ממייק**: חובה לאנשי NLP.

**בהירות כתיבה**: בינוני פלוס.

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

**יישומים פרקטיים אפשריים**: חקירה של תכונות מבניות של מודלי NLP במגוון סקאלות

---

**פרטי מאמר:**

**לינק למאמר**: זמין להורדה.

**לינק לקוד**: זמין כאן

**פורסם בתאריך**: 09.11.20, בארקיב.

**הוצג בכנס**: NeurIPS2020.

---

## תחומי מאמר:

- חקר תכונות מודלי NLP עמוקות.

## כלים מתמטיים, מושגים וסימונים:

- אנליזה ספקטרלית לגילוי של קשרים במגוון סקאלות בייצוג של טקסט (אמבדינג).
- התמרת קוסינוס דיסקרטית (DCT), התמרת קוסינוס דיסקרטית ההופכית (iDCT).
- מסנן מעביר נמוכים (LPF), מסנן מעביר גבוהים (HPF), מסנן מעביר פס (BPF).

---

## תמצית מאמר:

שפות טבעיות מתאפיינות בתכונות מבניות בכמה סקאלות שונות החל מהרמה של מילה עד רמת הפיסקה והמסמך. בהקשר זה נשאלת השאלה האם המודלים, המבוססים על רשתות הנוירונים בתחום NLP, תופסים את התכונות ההיררכיות אלו? האם ניתן "לשפר את ביצועי הרשת אם מאלצים אותה" לחקות את התכונות הללו? איך תכונות אלו משתנות בין מודלים, המאומנים למשימות שונות? המאמר הנסקר מנסה לתת מענה על השאלות האלו.

למעשה המאמר מציע שיטה לבחון תכונות וביצועי מודל NLP נתון בסקאלה נתונה עי" הורדותן של כל הסקאלות האחרות מהמודל. למשל בשביל לבדוק את ביצועי המודל בסקאלת קצרת טווח (רמת מילה) למשימה ספציפית, הם מאלצים את המודל "לא להשתמש" בסקאלות ארוכות טווח (משפטים, פסקאות וכדומה). זה נעשה עי" שימוש בטכניקות ספקטרליות מתחום עיבוד אותות המאפשרות לסנן (בתחום התדר) רק את התכונות בסקאלה הנדרשת. כאן סקאלות קצרות טווח (רמת מילה) מיוצגות עי" תדרים גבוהים כאשר סקאלות ארוכות טווח מיוצגות ע"י תדרים גבוהים יותר (נפרט על כך בהמשך).

השיטה המוצעת מסתמכת על הפעלה של מסננים ספקטרליים על אקטיבציות של נוירונים בשכבות שונות של הרשת לאורך הטקסט (זה מימד ה"זמן" שלנו !!). כלומר אם נרצה לבדוק עד כמה סקאלה קצרה (מילה או שתיים, תדרים גבוהים) משפיעה על ביצועי מודל, מוסיפים למודל שכבה המפלטרת החוצה את כל הסקאלות הארוכות (תדרים יותר נמוכים). אם ביצועי מודל לא משתנים בצורה משמעותית כתוצאה מסינון זה, המסקנה היא ש"תלויות (סקאלות) ברמת מילה" חשובות חשובות יותר לביצוע מוצלח של המשימה מאשר תלויות ארוכות טווח. כלומר במשימה זו "למודל מספיק להתמקד בתלויות קצרות טווח בטקסט" בשביל להשיג ביצועים טובים.

טכניקה זו מאפשרת לבודד את התכונות (מידע) הקשורות לסקאלה ולהפריד אותן מהתכונות הסמנטיות של וקטורי ייצוג של טוקנים. בשביל להגיע להפרדה זו מוסיפים למודל שכבה המעבירה חלקים שונים של וקטורי ייצוג של הטוקנים (אמבדינגס) דרך מסננים ספקטרליים שונים.

**הערה:** המאמר טוען שבעיקרון ניתן להוסיף שכבה מסננת (שנקראת Prism) לא רק בתור השכבה האחרונה של הרשת, אך בפועל בכל הניסויים שהם עשו, הם הוסיפו את Prism אחרי שכבת האמבדינגס של BERT. בעקבות זה אתייחס בהמשך רק לסינון הספקטרלי של שכבת ייצוג הטוקנים (אמבדינגס).

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

## הסבר של רעיונות בסיסיים:

בואו ננסה להבין איך בעצם עובדת שכבת Prism:

- חלוקה לסקאלות (תדרים): מחלקים את הרכיבים של וקטורי הייצוג לכמה תת-קבוצות. למשל אם יש לנו אמבדינגס באורך 360 ואנו רוצים לבחון 3 סקאלות שונות, הרכיבים 1,..., 120 (קבוצת אינדקסים S_1) יהיו "אחראים" על הסקאלה ראשונה עם התדרים הגבוהים ביותר (ברמת מילה עד שתי מילים נגיד), הרכיבים 121,..., 240 (קבוצה S_2) ייצגו את הסקאלה השניה עם התדרים הבינוניים (ברמת "המשפט") , ו- 120 הרכיבים האחרונים S_3 "ישויכו" לסקאלה 3 של התדרים הנמוכים ביותר (ברמת "פסקה/המסמך").
- בנייה של וקטורי דגימות T לכל נוירון באמבדינג: לכל אינדקס i בווקטורי הייצוג על פני כל הטוקנים בטקסט, בונים וקטור דגימות T_i. למשל עבור רכיב מסוים בווקטור הייצוג (נגיד במיקום 213) ובונים וקטור דגימות T_213 המורכב מכל הרכיבים מס' 213 על פני כל הייצוגים של הטוקנים בטקסט.
- העברה של וקטורי T_i דרך DCT: מפעילים את התמרת קוסינוס דיסקרטיות DCT (יפורט בהמשך) על כל וקטור D_i ובונים להם את הייצוגים הספקטרליים (בתחום התדר). הייצוג הספקטרלי של וקטור דגימות T_i יסומן ב F_i. נציין כי כל וקטורי דגימות עוברים אותה התמרה כלומר אם יש לנו 150 וקטורי T_i, אנו צריכים לבצע 150 DCTים (לכל אחד בנפרד). חשוב לזכור שהמימד של כל וקטור T_i שווה למספר הטוקנים בטקסט(!!).
- סינון ספקטרלי של וקטורי F_i: לכל וקטור F_i בוחרים את המסנן הספקטרלי שלו לפי האינדקס i. וקטורי F_i עם אינדקסים מקבוצה S_1 (ברמת מילה) יועברו דרך מסנן מעביר גבוהים HPF, האינדקסים מקבוצה S3 יועברו דרך מסנן מעביר נמוכים ואינדקסים מקבוצה S2 יועברו דרך מסנן מעביר פס BPF (ההסבר על איך עובדים המסננים נמצא בפרק הבא).
- העברה של וקטורי F_i המסוננים דרך התמרת קוסינוס ההופכית iDCT: למעשה iDCT מעבירה את הספקטרום המסונן של הייצוגים בחזרה לתחום "זמן" (נזכיר שאצלנו מימד הזמן זה האינדקסים של האמבדינגס לאורך הטקסט). נסמן את התוצאה של פעולה זו כ T_fi. שעבור כל i הווקטור T_fi בנוי מכל הרכיבים במיקום i של וקטורי הייצוג המסוננים.
- אימון רגיל של רשת (BERT) עם שכבת prism.

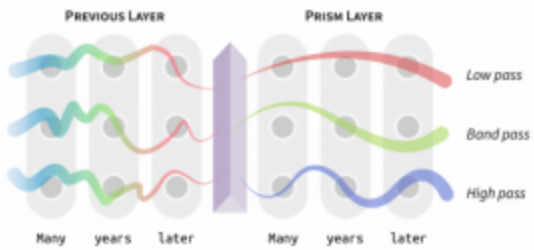


Figure 1: **The prism layer specializes different neurons for different scales.** First, the representations for an input are computed (left; in this case, the input is of length three). Next, a spectral filter (a low-, high-, or band-pass) is applied along the activations of each individual neuron (right). 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 "סקאלה של מילה" מייצגת דווקא תדרים גבוהים בזמן שה"סקאלה של מסמך" מייצגת דווקא את התדרים הנמוכים ביותר? התשובה לכך נובעת מהעובדה ש"התדר של סקאלה בטקסט" הינו ביחס הפוך ל"מחזור" של אותה סקאלה. למעשה "המחזור" של "מילה" הינו נמוך ביותר בזמן של מחזור של "סקאלת הפיסקה" הינו גבוה הרבה יותר. הסיבה לכך שהטקסט מורכב מהרבה מילים, פחות משפטים ועוד פחות פסקאות.

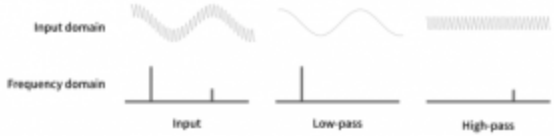


Figure 2: **A visual depiction of spectral filters and their effects in the input and frequency domain.** The input domain shows a sequence of values (e.g., the activation of a neuron across input tokens). The frequency domain shows the weight on the cosine waves which sum to produce the curve in the 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 סקאלות לביצועי מודל עבור משימות שונות: בשביל לבדוק את רמת ההשפעה של "סקאלה" מסוימת" על ביצועי המודל, המחברים סיננו את כל הסקאלות האחרות. נניח שאנו רוצים לבחון את ההשפעה של סקאלת "המילים" (תדרים גבוהים) על ביצועי מודל במשימה מסוימת. אז מפעילים מסנן שמסנן את כל התדרים האחרים (הנמוכים והבינוניים) על ידי העברה של ייצוגי הטוקנים לאורך הטקסט דרך HPF בצורה המפורטת בסעיף הקודם. המאמר חילק את הסקאלות (תדרים) ל-5 תחומים השווים באורך:

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

מהבדיקות המוצגות במאמר עולה כי למשימת זיהוי נושא, התדרים הנמוכים הם הכי חשובים שזה די הגיוני כי המודל צריך "להבין" את כך הטקסט כולו פחות או יותר בשביל לזהות את הנושא שלו. מה שקצת מפתיע בתוצאות שלהם זה השיפור המשמעותי בביצועים של המודל מול המודל המקורי אחרי 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 חלקי דיבור התדרים הגבוהים יצאו הכי משמעותיים שזה די מובן בהתחשב לאופי המשימה. הרי בשביל להבין לאיזה חלק דיבור לשייך מילה, מספיק לקחת בחשבון מילה או שתיים סמוכות.

מעניין שלמשימת זיהוי מילה ממוסכת שעליה אומן BERT (בנוסף לזיהוי סדר המשפטים) התדרים הכי חשובים הם הגבוהים ביותר כלומר בשביל לנחש מילה "תחת מסכה" מספיק לדעת מילה או שתיים מסביב אליה. בעיני זו תגלית מאוד מסקרנת(!!).

### ביצועי מודל עם שכבת Prism:

המחברים הוסיפו 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 עבור משימת זיהוי חלקי דיבור הם קיבלו תוצאות נמוכות טיפה מ-BERT המקורי. האימון בוצע על דאטהסט WikiText-103.

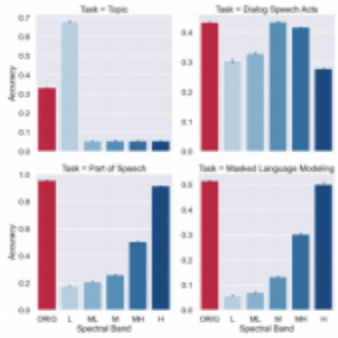


Figure 4: **Different spectral filters extract information useful for tasks at different scales.** Probing accuracy for different tasks and band-passes. A low-pass filter produces representations that yield highest probing accuracy on topic classification, while high-passed representations have highest 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 קוסינוס דיסקרטית DCT והופכית שלה IDCT: למעשה זה מקרה פרטי של התמרת פוריה הסטנדרטית. היא פועלת על סדרה של מספרים ממשיים ומעבירה אותה לסדרה ממשית מאותו אורך בתחום התדר. אינטואיטיבית, התמרה זו מחפשת דמיון בין הסדרה לפונקציות קוסינוס מתדרים שונים.

## דאטהסטים ומשימות:

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

## נ.ב.

מאמר עם תוצאות מאוד מסקרונות, המשתמש בטכניקות ספקטרליות לבחינה של תבניות (אורכי תלויות) עבור מודלי NLP עמוקים במשימות שונות. לצערי ביצועי הגישה המוצעת במאמר נבדקו על מעט משימות ורק על דאטהסט אחד בלבד לכל משימה. עובדה זו קצת מקשה עליי להשתכנע שהתופעות שהם גילו מתרחשים במשימות NLP אחרות בדאטהסטים אחרים. אני מצפה שהמשך של המחקר המעניין הזה...

#deepnightlearners

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

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