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:

פינת הסוקר:

המלצת קריאה ממייק ומעידו: מומלץ מאוד למי שרוצה להבין האם יש קשר בין משוואות דיפרנציאליות חלקיות (PDE) ובין רשתות נוירונים גרפיות (GNN)

בהירות כתיבה: בינונית פלוס

ידע מוקדם: הבנה בסיסית ב-PDE וב-GNN ובנוסף היכרות עם מושגי יסוד בתורת הגרפים במשוואות דיפרנציאליות.

יישומים פרקטיים אפשריים: פתרון של בעיות בביולוגיה חישובית, עיבוד תמונה, גרפיקה, ראיה, ניתוח רשתות חברתיות ועוד.

פרטי מאמר:

לינק למאמר: זמין להורדה.

לינק לקוד: ---

פורסם בתאריך: 26.10.21, בארקיב.

הוצג בכנס: NeurIPS 2021.

מקום השתייכות של המחבר הראשון: אוניברסיטת בן-גוריון בנגב.

תחומי מאמר:

- רשתות נוירונים על גרפים (GNN)
- אנליזה נומרית של משוואות דיפרנציאליות חלקיות
- פיזיקה חישובית (computational physics)

כלים מתמטיים, מושגים וסימונים:

- רשתות קונבולוציה על גרפים
- החלקת יתר (over-smoothing) ברשתות נוירונים על גרפים
- אופרטורים דיפרנציאליים כמו דיברגנץ, לפלסיאן
- רשתות קונבולוציה קלאסיות (CNNs)

תמצית מאמר:

אחת הסוגיות המרכזיות ברשתות קונבולוציה על גרפים (GCN) הינה החלקת יתר (over-smoothing) של ייצוג דאטה המופק באמצעות הרשת. החלקת יתר של ייצוג מאופיינת בשינויים הולכים וקטנים של בין ייצוג של פיסות דאטה שונות (כגון embeddings של קודקודים וקשתות) המופק באמצעות GCN. בעיה זו מחריפה בשכבות העמוקות של GCN כלומר ייצוגי קודקודים וקשתות בשכבות אלו נהיים זהים זה לזה. תופעה זו היא הסיבה העיקרית לכך שכותבי מאמרים בתחום ה-GCN נוטים להסתפק בכמות קטנה של שכבות לעומת רשתות הקונבולוציה הקלאסיות (CNNs). חולשה נוספת של GCNs היא הצורך בהתאמת הארכיטקטורה שלהן לדומיין ולמשימה. לדוגמה, כותבי המאמר PDE-GCN מציינים כי GCN, המבצעת משימת סיווג בענן נקודות, עלולה להפגין ביצועים ירודים במשימת סיווג של קודקוד בגרף ציטוטי מאמרים (citation network).

המאמר המסוקר מציע גישה מעניינת ומקורית הבא לתת מענה לסוגיות אלו. השיטה המוצעת מנצלת את הקשר הקיים בין רשתות נוירונים לבין משוואות דיפרנציאליות חלקיות (PDE). קשר זה נחקר בצורה אינטנסיבית בתקופה האחרונה (1, 2, 3, 4, 5). הדינמיקה של מפות הפיצ'רים לאורך שכבות של CNN ניתנת לתיאור באמצעות מערכת דינמית המתוארת על ידי משוואה דיפרנציאלית חלקית. כלומר ניתן להתייחס לכל שכבה של CNN כאל "צעד בזמן" של הגרסה הדיסקרטית של משוואה דיפרנציאלית חלקית. זהו משפט מאוד אבסטרקטי שקושר שני דברים שאנחנו לא רגילים לחשוב עליהם בעת ובעונה אחת, ולכן סקירה זו תנסה להתיר במקצת את הקשר הסבוך שעליו בונים מחברי PDE-GCN את התיאוריה במאמר.

הסקירה תחולק לשלושת החלקים הבאים:

- רקע על ההקבלה בין מד"ח לרשתות קונבולוציה

- מבוא מזורז לרשתות קונבולוציה על גרפים
- החיבור ביניהם והחידוש העיקרי של המאמר.

חומר רקע:

הקשר בין רשתות למשוואות דיפרנציאליות:

ננסה כעת להסביר (קצת בנפנופי ידיים - אם אתם רוצים הסבר ריגורוזי תעיפו מבט PDEs and Convolutions או באחד המאמרים המקושרים בפסקה הקודמת) איך PDEs קשורים ל-CNNs. נתחיל מרשת נוירונים די בסיסית ונראה איך ניתן "להפוך" אותה ל-PDE באמצעות מניפולציות די פשוטות. נניח שיש לנו רשת המורכבת מ-T שכבות בעלות חיבור שיורי (residual connection). העיקרון שנדון בו כאן רלוונטי לכל CNN, אבל ההסבר בהיר יותר כאשר מתייחסים אל ארכיטקטורה מבוססת ResNet, ובכל מקרה זו ארכיטקטורה סטנדרטית של רשת קונבולוציות.

כאמור, פלט x_{t+1} של שכבה $t = 1, \dots, T$ ניתן לתיאור[1] כ:

$$x_{t+1} = x_t + W_2\sigma(W_1 x_t),$$

כאשר W_1, W_2 הן מטריצות קונבולוציה, ו-σ הינה פונקציית אקטיבציה לא לינארית כמו סיגמואיד או ReLU. כעת, אם נעביר איבר אחד שמאלה נקבל את המשוואה הבאה:

$$x_{t+1} - x_t = W_2\sigma(W_1 x_t),$$

ואם נכליל את המשוואה הזו ונחליף את המקדם 1 של צד ימין בקבוע $h \leq 1$:

$$x_{t+1} - x_t = hW_2\sigma(W_1 x_t)$$

$$\frac{x_{t+1} - x_t}{h} = W_2\sigma(W_1 x_t)$$

אז צד שמאל של המשוואה האחרונה מזכיר את הקירוב מסדר ראשון של $\frac{\partial x}{\partial t}$ בשיטת ההפרשים הסופיים (finite differences). זהו פתח ראשוני אל עולם הקשרים בין מד"ח (או מישדיפ, אם תתעקשו 😉) ללמידה עמוקה.

כעת נותר לנו רק למצוא אנלוגיה לנגזרות המרחביות, שבלעדיהן אף משוואה דיפרנציאלית אינה מעניינת. לצורך כך, נתחיל מהכיוון השני ונתבונן תחילה במשוואת הדיפוזיה הקלאסית:

---

[1] ב-ResNet אמיתי, על xt בצד ימין יופעל בדרך כלל אופרטור downsample כלשהו כדי להתאים את המימדים שלו למימדי השכבה, אבל לצורך הדיון זה פרט שולי. כמו כן נתעלם כאן מאיברי bias שמופיעים לעיתים קרובות.

$$\frac{\partial x}{\partial t} = W\nabla^2 x = W\nabla \cdot (\nabla x)$$

כאשר W הינה מטריצה סימטרית וחיובית-לחלוטין (positive definite) כלשהי, ∇^2 הוא אופרטור הלפלסיאן ו-∇ ,$\nabla \cdot$ הם הדיברגנץ[2] והגרדיאנט[3], בהתאמה. נעניק לכל אחד מאיברי המשוואה את הגרסה הדיסקרטית שלו ונקבל:

$$\frac{x_{t+1} - x_t}{h} = WG^T G x_t$$

וכדי לראות את הקשר למשוואת הרזנט (ResNet) שלנו, נכליל גם את משוואה זו ונגדיר $W_1 = G,\ W_2 = WG^T$ כמקרה פרטי. אם נוסיף את פונקציית האקטיבציה במקום הנכון, קיבלנו בדיוק את משוואת הרזנט.

לא ניכנס כאן לפרטים הטכניים של האנלוגיה הזו ונסתפק בנפנופי ידיים, אבל על בסיס זה מרשה לעצמו המאמר להתייחס לרשת CNN בתור הכללה של משוואת דיפוזיה, כאשר הקונבולוציות משחקות תפקיד של "אופרטורים דיפרנציאליים" נלמדים אשר מותאמים לדאטה שעליו מתאמנת הרשת ע"י תהליך האימון.

יסודות של GCN:

רשת נוירונים לגרף מגיעה מהצורך לעבד סיגנלים שחיים בעולם לא-סדור (unstructured). ניתן לראות כי גרף הוא הכללה של תמונה, בו הפיקסלים במרווחים לא קבועים ומספר השכנים משתנה גם הוא. כך ניתן לראות רשת נוירונים על גרף בתור הכללה של רשת נוירונים סטנדרטית המותאמת לסוגי דאטה לא-סדורים. הפעם אנחנו מתעניינים ברשת Graph Convolutional Network - GCN, רשת אשר פועלת על גרף באמצעות קונבולוציות.

המטרה של רשת כזו היא ללמוד ייצוגים של הקודקודים והקשתות בגרף באמצעות העברת אינפורמציה מקודקודים וקשתות אחרים בגרף, בדרך כלל תוך התחשבות בקישוריות אשר נתונה לנו בגרף (ועם זאת, יש מודלים שממציאים קשתות או לומדים לחזות חלקים חדשים מהגרף, הכל לפי צורכי המשימה). כלומר, בהינתן ייצוג התחלתי לכל קודקוד וקשת בגרף, מעדכנים את הייצוג שלו על ידי הזרמת מידע מהקודקודים השכנים. העדכונים האלה בדרך כלל מתבססים על פילטרים ואופרטורים נלמדים, כמו ברשת CNN רגילה.

---

[2] תזכורת: אופרטור הדיברגנץ מוגדר במרחב אוקלידי כסכום הנגזרות החלקיות הכיווניות של שדה וקטורי או סקלרי.

[3] חשוב לשים לב: מדובר בגרדיאנט **מרחבי** ולא בגרדיאנט של פונקציית הלוס של הרשת, כפי שהתרגלנו. בהקבלה לרשת CNN שפועלת על תמונות - הגרדיאנט הזה יהיה הגרדיאנט (הדמיוני) של התמונה, כמו שמקובל לחשוב עליו בעיבוד תמונה קלאסי. דוגמאות ניתן למצוא בשפע בוויקיפדיה ובאינטרנט.

תקציר מאמר:

כמו שכבר אמרנו בתחילת הסקירה הגישה המוצעת באה להתמודד עם החלקת יתר של הפיצ'רים בין קודקודים וקשתות בשכבות העמוקות של GCN. המחברים מציעים ארכיטקטורה של GCN המבוססת על דיסקרטיזציה של **משוואה היפרבולית לא-לינארית** (כמובן, עם תנאי התחלה עבור $f(t, :)$ ותנאי שפה שלא נעסוק בהם כאן):

$$f_{tt} = \nabla \cdot K^{*}\sigma(K\nabla f)$$

המאמר מוכיח כי פתרון של משוואה זו לא גורם לשחיקה עבור ערכים גבוהים של t (נזכר כי t הוא למעשה מספר השכבה ב-GCN בגרסה הדיסקרטית של המשוואה). המאמר גם מראה כי PDE המתארת GCN סטנדרטי הינה **משוואת דיפוזיה לא-לינארית**:

$$f_{t} = \nabla \cdot K^{*}\sigma(K\nabla f)$$

למי שאינו מנוסה במד"ח והמשוואות נראות לו זהות - שימו לב לצד שמאל של שתי המשוואות. צד ימין אכן זהה. המחברים מראים כי הפתרון של המשוואה האחרונה מכיל מעט מאוד מידע מאחר ודינמיקת הערבוב של משוואת הדיפוזיה מיצעה את כל הפיצ'רים, והתופעה הזו חמורה יותר ככל שמספר השכבות גדל. לטענת המאמר זו הסיבה העיקרית לתופעת החלקת היתר המתרחשת בשכבות העמוקות של GCN סטנדרטיות.

כדי להתגבר על החלקת היתר הזו, המאמר מציע לבנות ארכיטקטורה חדשה של GCN הנקראת PDE-GCN, בהתבסס על **המשוואה ההיפרבולית ולא על משוואת הדיפוזיה כמו GCN סטנדרטי.** בנוסף המחברים מגדירים גירסה דיסקרטית של הגרדיאנט המרחבי $\mathbf{G}$ של הגרף: עבור שני קודקודים i ו-j שמחוברים בקשת, הגרדיאנט $\mathbf{G}_{ij}$ ביניהם מוגדר כהפרש של וקטורי הפיצ'רים (ייצוגי הקודקודים עבור השכבה הנוכחית) $\mathbf{f}_i$ ו-$\mathbf{f}_j$ המוכפלים במשקולות W_{ij} כלשהם (אשר יכולים להיות נלמדים או מהונדסים). נציין כי $\mathbf{G}$ הוא למעשה מיפוי ממרחב הקודקודים V למרחב הקשתות של הגרף E.

את אופרטור הדיברגנץ ($\nabla \cdot$), המופיע במשוואה ההיפרבולית ניתן לקרב באמצעות $\mathbf{G}^T$ (הדיברגנץ על גרף בדרך כלל מוגדר כמיפוי ממרחב הקשתות E למרחב הקודקודים V וכאן אנו פשוט משחלפים את $\mathbf{G}$ שהיא מיפוי מ- V ל-E) עכשיו רק נותר להפעיל את שני האופרטורים ברצף כדי לקבל את הביטוי $\mathbf{G}^T\mathbf{G}$ – עבור האגף הימני של המשוואה ההיפרבולית. אחרי שהגדרנו את כל המשתנים ניתן לבנות את הארכיטקטורה הבסיסית של שכבת PDE-GCN:

$$\mathbf{f}^{(t+1)} = 2\mathbf{f}^{(t)} - \mathbf{f}^{(t-1)} - h^2\mathbf{G}^T\mathbf{K}_t^T\sigma(\mathbf{K}_t\mathbf{G}\mathbf{f}^{(t)})$$

כאשר $\mathbf{K}_t$ היא מטריצת קונבולוציה 1x1 נלמדת ופונקציית האקטיבציה שנבחרה היא tanh.

אחרי שהגדרנו את המבנה של PDE-GCN נותר לנו להסביר איך כל העסק עובד בפועל. באמת שזה לא מסובך:

- לוקחים את הפיצ'רים של הקודקודים והקשתות,
- מעבירים אותם דרך שכבת קונבולוציות 1x1,
- מחשבים את הפלט של השכבות הבאות בהתבסס על המשוואה האחרונה.

הערה 1: המאמר גם מציע דרך לנצל פיצ׳רים על קשתות (אם הם זמינים) כקלט ל-PDE-GCN.

הערה 2: המאמר גם מציע שיטה לבניית ארכיטקטורה של PDE-GCN המבוססת על שילוב (צירוף קמור בגדול) של משוואות הדיפוזיה והמשוואה ההיפרבולית.

הישגי מאמר:

המאמר השווה את ביצועי PDE-GCN עבור ערכים שונים של מספר שכבות עם מגוון של GCN-ים על כמה משימות ודאטהסטים שונים. המטרה הייתה להראות כי הארכיטקטורה המוצעת מצליחה להתגבר על בעיית החלקת יתר של הפיצ'רים בשכבות העמוקות של GCN. הדרך הטבעית להוכיח זאת היא להראות כי לא נצפית ירידה בביצועי GCN כאשר מוסיפים לה שכבות (כמובן שאם החלקת יתר עדיין קיימת, הוספת שכבות לרשת עלול לפגוע בביצועים). המאמר אכן מראה כי בכל המשימות שנבחנו ביצועי PDE-GCN לא סופגים ירידה (אלא משתפרים קצת ברוב המקרים). בחלק מהמשימות המאמר אפילו מציג SOTA חדש, אך לא בכולן.

כאן צריך לציין שבהשוואה לרוב הארכיטקטורות האחרונות של GCN, הביצועים לא תמיד טובים יותר והרשתות מהשנים האחרונות יודעות להתמודד בהצלחה גם עם הרבה שכבות (במאמר נבדקו עד 64). המאמר למעשה מציג שיטה נוספת (ופשוטה יחסית) לתכנן רשתות GCN עמוקות (לעומת vanilla GCN של Kipf and Welling מ-2016 ו-Dropedge).

נ.ב.

המאמר מציע גישה חדשנית ומעניינת לבניית ארכיטקטורה של GCN באמצעות משוואות דיפרנציאליות חלקיות. הגישה מצליחה להתגבר על בעיית החלקת יתר המתרחשת כאשר מוסיפים שכבות לרשת. לדעתנו, זאת גישה מעניינת לתכנון ארכיטקטורות דיפ בכללי ו-GCN בפרט ואנו רואים במאמר זה (ובסקירה זו) פתח לתוך עולם מעניין, אך פחות מוכר. בראיה זו, סקירה היא הראשונה מבין מספר סקירות מאמרים בעולם האנלוגיות בין מד״ח ובין ארכיטקטורות דיפ.

#deepnightlearners

הפוסט נכתב על ידי מיכאל (מייק) ארליכסון, PhD, Michael Erlihson ועדו בן-יאיר.

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

עדו הוא סטודנט לתואר שני במדעי המחשב באוניברסיטת בן-גוריון בנגב, בתחום המדע החישובי ורשתות גרפים. עדו עבד בתחומים מגוונים כמו עיבוד תמונה וגרפיקה ממוחשבת בחברות כמו מיקרוסופט, וויקס ולייטריקס ויש לו ניסיון עשיר בהנדסת תוכנה.