המאמר היומי של מייק - 25.02.25  
Understanding Visual Feature Reliance through the Lens of Complexity

המאמר שאני סוקר היום מציג מחקר יוצא דופן, נדיר ומעניין על מורכבות פיצ'רים המופקים על ידי מודלים דיפ (אין RAG, סוכנים ו-LLMs שם :). מאמר זה קשור הדוקות לרעיון של צוואר הבקבוק של המידע ברשתות עצביות עמוקות, שטבע נפתלי תשבי.

המאמר מציג מסגרת תיאורטית-אינפורמציונית חדשה לכימות מורכבות פיצ'רים במודלי דיפו ומציע גישה מתמטית להבנה פיצ'רים, מתי והיכן פיצ'רים מופיעים במהלך האימון. בניגוד לשיטות מסורתיות שמתמקדות בסליינסי (saliency) ושיוך פיצ'רים (attribution), המחקר מציע את מידת המורכבות שקיבלה שם v-information כמדד למורכבות חישובית, אשר מבטא את המאמץ הנדרש כדי לחלץ פיצ'רים במקום לשערך רק את התלות הסטטיסטית הישירה שלה בקלט.

המחקר בוחן באופן שיטתי את התפתחותן בזמן אימון, התפלגותן המרחבית ותפקידן של פיצ'רים במודלים ויז'ן. הממצאים מצביעים על כך שמודלי דיפ מציגים תהליך למידה היררכי, שבו פיצ'רים פשוטים ודלות-מורכבות מופיעות מוקדם באימון ומתקדמות בקלות דרך חיבורים residual, בעוד פיצ'רים מורכבים יותר דורשות עיבוד עמוק יותר וזמן אימון ארוך יותר אך תורמות פחות משמעותית להחלטות הסופיות ממה שהיה מקובל להניח.

גישה מבוססת למורכבות בלמידת פיצ'רים

ניתוח פיצ'רים בלמידה עמוקה התמקד עד כה בעיקר בחישוב החשיבות והשימושיות שלהן למשימה כזו או אחרת, אך כמעט ולא בוצע ניסיון לכמת כמה מורכב לחלץ פיצ'ר מתוך דאטה. מחקר זה משנה את נקודת המבט המסורתית בכך שהוא מציע מדד למאמץ חישובי הדרוש ללמידת פיצ'ר.

הגדרה מחדש של מורכבות פיצ'רים

שיטות מסורתיות לשערוך פיצ'רים מסתמכות על שערוך מידע הדדי (mutual information) בין פיצ'ר לבין הדאטה. עם זאת, גישה זו אינה מביאה בחשבון את הקושי החישובי הכרוך בחילוץ הפיצ'ר.

החדשנות המרכזית במאמר היא ההצגה של v-information, המאפשר כימות של:

כמה עיבוד דרוש כדי לחלץ פיצ'ר מתוך שכבות הרשת.

עומק ועוצמת הטרנספורמציות הלא-לינאריות הנדרשות כדי להפיק את הפיצ'רי מקלט.

סיבוכיות מיפוי של קלט למרחב הפיצ'רי, במקום רק מדידת התלות הסטטיסטית שלהן בקלט.

מדוע מורכבות חישובית חשובה בלמידת פיצ'רים?

מחקרים קודמים תיאוריית צוואר הבקבוק האינפורמטיבי (Information Bottleneck Theory) מצביעים על כך שמודלים עמוקים מעבדים את הייצוגים באופן הדרגתי, תוך סינון מידע לא רלוונטי ושימור אותות משמעותיים למשימה. המחקר הזה מרחיב עקרונות אלו בכך שהוא מספק מדד כמותי להערכת אילו פיצ'רים דורשים עיבוד עמוק ואילו מופיעות כבר בשלבים מוקדמים יותר של הלמידה.

הנתונים האמפיריים תומכים בטענה שהמודלים מעדיפים ללמוד פיצ'רים פשוטים ודלות-מורכבות בשלבים המוקדמים של האימון, בעוד שבפיצ'רים מורכבים יותר מופיעות רק לאחר זמן אימון ממושך יותר. תוצאה זו עולה בקנה אחד עם תיאוריות הלמידה המדורגת (Curriculum Learning), לפיהן משטח אופטימיזציה של מודלים עמוקים נוטה לטובת למידת תבניות פשוטות תחילה לפני המעבר לאבסטרקציות מורכבות יותר.

דינמיקת הזמן של מורכבות פיצ'רים בזמן אימון

אחד הממצאים המרתקים ביותר במאמר הוא כי למידת הפיצ'רים מתרחשת בהדרגה על פני שלבי האימון:

שלבי האימון הראשונים: המודל לומד במהירות פיצ'רים דלי-מורכבות, אשר דורשות פחות טרנספורמציות לא-לינאריות.

שלבי האימון האמצעיים: מתחילות להופיע פיצ'רים מורכבים יותר, המורכבות משילוב של תכונות פשוטות מוקדמות יותר.

שלבי האימון המאוחרים: פיצ'רים המורכבים ביותר מופיעים, אך תרומתן להחלטות המודל קטנה יחסית לעומת הפיצ'רים הראשוניים.

### התפלגות מורכבות הפיצ'רים במרחב הרשת הנוירונית

הממצאים מצביעים על כך שמורכבות פיצ'רים אינה מפוזרת באופן אחיד על פני שכבות הרשת, אלא מאורגנת בצורה מבנית:

פיצ'רים פשוטים מופיעות בשכבות המוקדמות ויכולות להתקדם דרך חיבורי residual.

פיצ'רים מורכבים דורשות עיבוד עמוק יותר ומצטברות בהדרגה דרך טרנספורמציות לא-לינאריות רבות.

חיבורי residual משמשים כמסננים חישוביים, ומאפשרים לפיצ'רים דלי-מורכבות לעקוף עיבוד עמוק שאינו הכרחי עבורן.

### הקשר בין מורכבות הפיצ'רים להחלטות המודל

אחד הממצאים המעניינים של המחקר הוא שפיצ'רים מורכבים משפיעים פחות על החלטות הסיווג הסופיות של המודל מאשר פיצ'רים פשוטים יותר.

מודלים מסתמכים בעיקר על פיצ'רים פשוטים ויציבות לצורך הכללה.

פיצ'רים מורכבים, על אף שהן קיימות, אינן חיוניות להכרעת הסיווג.

התמקדות יתר בפיצ'רים מורכבים אינה משפרת בהכרח את הביצועים, ועלולה להוביל לאוורפיט.

ממצא זה סותר את ההנחה המסורתית שלפיה מודלים עמוקים מסתמכים בעיקר על ייצוגים אבסטרקטיים מאוד לשם קבלת החלטות. המאמר מציע כי המודלים מנצלים קודם כל פיצ'רים פשוטים ועמידים, ורק אחר כך משלבים מידע מורכב יותר כתוספת רפינמנט משנית.

## סיכום:

המאמר מציע תרומה תיאורטית ואמפירית משמעותית להבנת כיצד מודלים עמוקים לומדים, מארגנים ומשתמשים בפיצ'רים שונות. תובנות אלו יכולות להשפיע על עיצוב ארכיטקטורות רשת, אסטרטגיות אימון, ופרשנות של למידת מכונה, תוך שיפור היעילות והעמידות של מערכות בינה מלאכותית.

https://arxiv.org/abs/2407.06076