Deep k-Nearest Neighbors - Towards Confident, Interpretable and Robust Deep Learning

הרעיון המרכזי

המאמר מציע שיטה בשם Deep k-Nearest Neighbors, או DkNN.

במקום לסמוך רק על ה - softmax הסופי של רשת עמוקה, החוקרים בודקים האם התחזית של המודל נתמכת גם על ידי הייצוגים הפנימיים שלו.
כלומר, עבור קלט חדש, הם בודקים בכל שכבה של המודל - לאילו דוגמאות אימון הקלט הזה הכי דומה?

אם המודל חזה label מסוים, אבל השכנים הקרובים בשכבות הפנימיות שייכים ל - labels אחרים, אז התחזית נחשבת פחות אמינה.

הרעיון המרכזי - תחזית טובה היא לא רק תחזית עם softmax גבוה, אלא תחזית שיש לה תמיכה עקבית מתוך דוגמאות האימון לאורך שכבות המודל.

הבעיה שהמאמר מזהה

המאמר מתמודד עם בעיה מוכרת ב - Deep Learning - מודלים יכולים להיות מאוד בטוחים בעצמם גם כשהם טועים.

במיוחד במקרים של:

- קלטים מחוץ להתפלגות האימון
- דוגמאות adversarial
- קלטים עמומים
- הערה שלי (לידור) – עד כאן, נשמע כמו קלטים שבהם מודלי NLI נופלים.

כלומר - High softmax confidence ≠ reliable prediction

הבעיה אינה רק שהמודל טועה, אלא שהוא לא יודע לומר מתי אין לו באמת בסיס טוב לתחזית.

המתודולוגיה המוצעת

DkNN עובד כך :

1. מאמנים DNN רגיל.
2. לאחר האימון, שומרים עבור כל דוגמת אימון את הייצוגים שלה בכל שכבה.
3. בזמן inference, מעבירים קלט חדש דרך המודל.
4. בכל שכבה מחפשים את k השכנים הקרובים ביותר מתוך דוגמאות האימון.
5. בודקים את ה - labels של השכנים.
6. אם השכנים לאורך השכבות תומכים בתחזית הסופית - התחזית נחשבת אמינה יותר.
7. אם יש חוסר התאמה בין התחזית הסופית לבין השכנים הפנימיים - התחזית נחשבת חשודה.

איך בפועל נבדקת התמיכה בתחזית

חשוב להבין ש - DkNN לא מחפש שכנים "ללייבל", אלא שכנים לייצוג של הקלט בכל שכבה.
כלומר, לאחר שמעבירים קלט חדש ברשת – מבצעים את אלגוריתם KNN על הייצוגים הפנימיים של השכבות של הרשת, בכל שכבה בנפרד.
במילים אחרות, עבור כל שכבה i של הרשת, יש i מודלי KNN, כך שכל מודל בודק את K הייצוגים הקרובים ביותר לייצוג של שכבה i.

רק אחרי שמוצאים את K השכנים בכל שכבה, מסתכלים על ה - labels שלהם.
אם המודל חזה label מסוים, ורוב השכנים לאורך השכבות שייכים לאותו label, אז התחזית מקבלת תמיכה חזקה מדוגמאות האימון.
לעומת זאת, אם בשכבות שונות השכנים מגיעים מ - labels אחרים, אז יש סימן שהתחזית פחות יציבה או פחות נתמכת.

Nonconformity ו - Credibility

Nonconformity הוא מדד האומר כמה שכנים, מכל השכבות יחד, לא תומכים ב - label מסוים.

לדוגמה:
אם המודל חזה "7", אבל מתוך כל השכנים שנמצאו בשכבות השונות הרבה מהם הם בכלל "2", "3" או "9", אז ה - nonconformity של "7" יהיה גבוה.

כלומר:

- nonconformity גבוה = מעט תמיכה מהשכנים.
- nonconformity נמוך = הרבה תמיכה מהשכנים.

אבל כדי לדעת האם זה באמת "גבוה", המאמר לא קובע סף ידני. במקום זאת, הוא משווה את המצב לדוגמאות רגילות מתוך calibration set.

מכאן מתקבל מדד ה - **credibility**:

- Credibility גבוהה = התחזית נראית כמו תחזית רגילה שנתמכת היטב בדוגמאות האימון.
- Credibility נמוכה = המודל נתן תחזית, אבל אין לה תמיכה חזקה מספיק מהשכנים הפנימיים.

ה - calibration set משמש כנקודת ייחוס. מחשבים עליו איך נראית רמת nonconformity רגילה, ואז משווים אליה את הדוגמה החדשה. לכן credibility נמוכה לא נקבעת לפי סף ידני של מספר שכנים, אלא לפי השאלה האם חוסר-התמיכה בתחזית חריג ביחס לדוגמאות תקינות שהמודל אמור להכיר.

במילים פשוטות - DkNN לא שואל רק "מה המודל חזה?", אלא גם "האם יש לדאטה האימוני ראיות שתומכות בתחזית הזו?"

המודלים / הדאטה שנבדקו

המאמר לא בודק מודלי NLI או מודלי שפה. הוא בודק רשתות ראייה ממוחשבת על datasets כמו:

- MNIST
- SVHN
- GTSRB

בנוסף, יש דוגמה על ResNet ו - ImageNet לצורך interpretability.

כלומר, זה מאמר מעולם של - Deep Learning ו - Computer Vision, ולא מאמר NLP / NLI.

התוצאות המרכזיות

המאמר מראה ש - DkNN כמעט לא פוגע ב - accuracy הרגיל, אבל מוסיף יכולת טובה יותר לזהות תחזיות לא אמינות.

הממצא החשוב - כאשר הקלט הוא out-of-distribution או adversarial, softmax עדיין נוטה לתת confidence גבוה, **אבל** DkNN נותן credibility נמוכה.

כלומר, DkNN טוב יותר בלומר - "המודל נתן תחזית, אבל אין לה תמיכה חזקה בדאטה."

בנוסף, במקרים מסוימים DkNN משפר עמידות מול adversarial examples, כי הוא מזהה שהתחזית הסופית לא נתמכת לאורך השכבות הפנימיות.