# Review 183: [Short] WEAK-TO-STRONG GENERALIZATION: ELICITING STRONG CAPABILITIES WITH WEAK SUPERVISION

**Paper: https://arxiv.org/abs/2312.09390v1**

https://openai.com/research/weak-to-strong-generalization, https://cdn.openai.com/papers/weak-to-strong-generalization.pdf

אוקיי, הסקירה של היום על מחקר מאוד מעניין מבית לא אחר אלא OpenAI. המאמר מנסה לפתח שיטות אימון (כיול) למודלים חזקים עם מודל פיקוח חלש. מה זה בעצם אומר? נניח שיש לנו מודל S בעל יכולות חזקות יותר (נגיד במספר פרמטרים) ממודל חלש יותר W ובנוסף יש לנו דאטהסט לא מתויג שאנחנו רוצים לכייל על מודל S.

פיקוח חלש אומר שאנו קודם כל מתייגים את הדאטה עם W ואז מאמנים מודל חזק S עם הדאטהסט המתויג הזה. למה זה בעצם חשוב? המאמר מדבר בעתיד הקרוב יחסית אנו נגיע לאימון של מודלים בעלי יכולות superhuman למשימות שאנו בני אדם לא מסוגלים לבצע באיכות טובה ואז אנחנו בעצם מהווים את המודל החלש W.

השאלה עד כמה אימון מודל בפיקוח חלש עובד גרוע יחסית לכיול של מודל S עם דאטה מתויג נכון (על ידי בני אדם)? מתברר שההפרש בביצועים הוא די גדול למרות שאימון בפיקוח חלש כן מצליח לשפר את הביצועים של המודל החזק.

השאלה בעצם האם יש שיטות שמשיגות ביצועים טובים יותר מאשר אימון בפיקוח חלש? המאמר מציע שתי שתי משפחות של שיטות. השיטה הראשונה היא אימון הדרגתי של המודל החזק. מתחילים ממודל חלש וכל פעם ״מחזקים״ אות בקצת (בכמות הפרמטרים למשל) כאשר המודל מהאיטרציה הקודמת משרת בתור מודל פיקוח חלש. הגישה השנייה (consistency loss)היא לתת פחות לקנוס את המודל החזק על אי התאמה עם המודל החלש כאשר המודל החזק מאוד בטוח בתוצאה שלו.

יש תוצאות מעניינות…