# Review 26: What does a platypus look like? Generating customized prompts for zero-shot image classification

**Paper: https://arxiv.org/abs/2209.03320v3**

פינת הסוקר:

המלצת קריאה ממייק וליאור: מומלץ למי שמתעניין במשימות סיווג דאטה ויזואלי עם מילון פתוח (של קטגוריות)

בהירות כתיבה: בינונית.

ידע מוקדם:

הבנה בעקרונות של מודלי שפה

הבנה בסיסית בפרומפטים

ידע במודלים cross-domain כמו CLIP ובסיווג עם מילון פתוח

יישומים פרקטיים אפשריים:

יצירה מהירה של פרומפטים למשימות סיווג בעלות מילון קטגוריות פתוח

פרטי מאמר:

פורסם בתאריך: 07.09.2022, בארקיב.

הוצג בכנס: ---.

תחומי מאמר:

סיווג תמונות באמצעות מילון פתוח (open-vocabulary classification)

כלים מתמטיים, מושגים וסימונים:

מודלי שפה גדולים (LLM) כמו GPT3

מודלים cross-domain כמו CLIP המאפשרים סיווג zero-shot של תמונות

מבוא:

מטרת משימת סיווג תמונות היא לאמן מודל לזיהוי הקטגוריה של תמונה נתונה. בעבר משימה זו הניחה קיום של מילון קטגוריות קבוע מראש ולא משתנה במהלך אימון המודל (למשל 1000 קטגוריות של ImageNet). בשנה האחרונות אנו רואים יותר ויותר מאמרים שמשתמשים בגישה מבטיחה לסיווג תמונות שהיא מבוססת על מילון קטגוריות פתוח (open vocabulary - OVC classification). בשונה משיטות סיווג ישנות יותר OVC מאפשרת סיווג תמונות עם קבוצה שרירותית ולא סגורה של קטגוריות המוגדרות באמצעות שפה טבעית. למשל תמונה של דוב יכולה להיות מסווגת עם מודל OVC בתור ״דוב שחור אמריקאי שנפוץ באמריקה הצפונית״.

מודלי OVC השיגו ביצועים גבוהים על משימות סיווג על מספר רב של דאטהסטים בהעדר דאטה מתויג (כלומר zero-shot). איך זה קרה, אתם שואלים? מודלים אלה ממנפים כמויות אדירות של צמדי תמונה-טקסט (= תמונות עם כותרת) זמינים באינטרנט כדי למפות אותם (תמונות והתיאור שלהם) לאותו מרחב לטנטי. במרחב לטנטי זה השיכונים (embeddings) של תמונה והכותרת שלה נמצאים קרוב אחד לשני כאשר השיכונים של תמונה וטקסט לא קשור אליה ממוקמים רחוק אחד מהשני. הדוגמא המפורסמת ביותר של מודל כזה המנוצל בצורה אינטנסיבית במספר משימות ML היא CLIP.

מודלים כאלו מאפשרים לסווג תמונות על ידי הדמיון בין ייצוגי התמונה והכותרת. כדי לסווג תמונה נתונה יש צורך ביצירת פרומפטים (prompts) המתארים כל קטגוריה בצורה המיטבית. לאחר מכן הדמיון בין השיכון של הפרומפט הטוב ביותר לקטגוריה לבין שיכון התמונה יספק את מידת ההתאמה של קטגוריה זו לתמונה. באופן כזה נוכל לבצע סיווג ללא אימון נוסף וללא צורך בדאטהסט מתויג (כלומר zero-shot).

תמצית מאמר:

הגישה המתוארת הינה מאוד נוחה לשימוש ועשויה לחסוך לא מעט זמן אך היא טומנת בה מוקש קטן והוא מתחבא בבניית פרומפט המייצג קטגוריה נתונה בדרך הטובה ביותר. הגישה הסטנדרטית היא לייצר כמה תבניות גנריות לפרומפטים כמו "תמונה של {קטגוריה}״ ואז ליצור פרומפטים לכל קטגוריות של המשימה. לאחר מכן אנו מחשבים דמיון בין השיכונים של סט הפרומפטים לקטגוריה לבין השיכון של התמונה ומסווגים את התמונה לפי הקטגוריה עם סט הפרומפטים הכי דומה לה שיכון התמונה.

לשיטה זו יש כמה חסרונות משמעתיים. ראשית, כל פרומפט נבנה בצורה ידנית וזה מצריך בנייה מחדש של פרומפטים עבור כל דאטהסט חדש שהופך פעולה זו ללא סקיילבילי. שנית, הפרומפטים חייבים להיות כלליים מספיק כדי לחול על כל קטגוריות התמונות. לדוגמה, פרומפט חייב להיות כללי מספיק: ב-ImageNet הקטגוריה של פלטיפוס (platypus) יכול להיות ״תמונה של {פלטיפוס}״ ולא יכולה להיות משהו כמו ״תמונה של {פלטיפוס}, יונק מימי״ כי זה לא יתאים לא קטגוריות אחרות. לבסוף, כתיבת פרומפטים איכותיים עם דורשת ידע על הדאטהסט עצמו שלא ניתן להכלילו על דאטהסטים אחרים וזה מאריך את הזמן הנדרש לבניית מודל.

המאמר מציע גישה חדשה לבניית פרומפטים שנותנת מענה לאתגרים המתוארים בפסקה הקודמת. במקום לבנות פרומפטים בצורה ידנית החוקרים מנצלים מודל שפה גדול לבנייתם. ההתערבות האנושית נדרשת רק בשלב הראשון לבנייה של כמה פרומפטים כלליים בודדים כמו ״איך נראה פלטיפוס?״ וכדומה. למרות שזה דורש קצת ״הנדסה ידנית״, זה מתגמד מול כמות פרומפטים שצריך לייצר בשיטה הסטנדרטית (המתוארת לעיל). פרומפטים כאלו נקראים פרומפטי LLM במאמר. כדי ליצור פרומפטים שניתן לנצל אותם למשימת OVC, מזינים פרומפטים אלו למודל שפה מאומן (המאמר משתמש ב-GPT3 למשימה זו). פרומפטים כאלו נקראים פרומפטי התמונה (image prompts).

פרטים נוספים של הגישה המוצעת:

המחברים ניסו את שני הוריאנטים הבאים של גישה זו:

CuPL Single Prompt: כתיבה של פרומפט אחד לכל הדאטהסטים לפי הטמפלייט הבא:  
  
 "Describe what a(n) {category}, a type of \_, looks like"

החלק המודגש מכיל סוג כללי שהקטגוריה שייכת אליו (כמו חיית בית או מטוס לדאטהסטים כמו Oxford Pets או FGVC Aircraft). לדאטהסטים כלליים יותר כמו ImageNet חלק זה נמחק מהפרומפט.

CuPL Full Prompts - שימוש בפרומפטי LLM שונים לכל דאטהסט (בסך הכל 5 לכל דאטהסט). למשל ל-ImageNet הפרומפטים הבאים נוצרו:

"Describe what a(n) {} looks like",

“How can you identify a(n) {}?"

"What does a(n) {} look like?"

“A caption of an image of a(n) {}:"

"Describe an image from the internet of a(n) {}”

המחברים מציינים שהווריאנט השני דורש יותר מאמץ ליצירה של פרומפטים אלו אך עדיין בכמות נמוכה משמעותית ממה שנדרש בשיטות הקודמות.

לאחר היצירה של פרומפטי LLM לכל קטגוריה, הם מוזנים למודל שפה גדול (GPT3) והפלט שלו משמש ליצירה של פרומפטי תמונה כאשר עבור כל פרומפט LLM נוצרים 10 פרומפטי תמונה. לאחר מכן מחשבים את השיכון (embedding) של פרומפטי התמונה שנוצרו לכל קטגוריה וממצעים אותם (לכל קטגוריה בנפרד) כדי לבנות ייצוג של כל קטגוריה. בשלב האחרון מחשבים את הדמיון בין השיכונים של הקטגוריות עם שיכון התמונה ומסווגים את התמונה עם הקטגוריה בעלת דמיון הגבוה ביותר.

הישגי המאמר:

החוקרים ערכו ניסוי לבדיקת ביצועי הסיווג של שני הווריאנטים לגישה המוצעת ליצירת פרומפטים, ביחס לשיטה הסטנדרטית אשר משמשת כ-baseline. השיטות נבדקו על שמונה benchmarks מוכרים של סיווג תמונות בשיטת Zero Shot.

מעבר לתוצאות הסיווג נבדקו גם ההשפעות של גודל המודל, כמות הפרומפטים והמגוון שלהם על הביצועים.

על מנת לממש OVC נדרש שכל קטגוריית תמונות תיוצג בשפה טבעית בצורה חד משמעית. לא בכל הדאטהסטים הדבר זמין. כך למשל, ב-ImageNet הקטגוריות מיוצגות כ-ID אשר יכול להיות ממופה למספר מילים נרדפות ב-WordNet. לכן, בעבודה זו הם השתמשו בתוויות כפי שהוגדרו בעבודה קודמת בנושא.

תוצאות הניסוי:

עבור כל דאטהסט מוצגים ה-top 1 accuracy של הסיווג, הדלתא (ההפרש) ביחס לשיטה הסטנדרטית (ירוק = שיפור), ומספר הפרומפטים הידניים שנוצרו.

יש לשים לב שבמקרה של השיטה הסטנדרטית מספר הפרומפטים הידניים מתייחס לפרומפטי התמונה, ובשיטות CuPL המספר מתייחס לפרומפטי ה-LLM.

כאמור בשיטת CuPL Single, נדרש פרומפט יחיד לכל הדאטהסים, לעומת 97 סה"כ, בשיטה הסטנדרטית. מעבר לכך, הפרומפטים בשיטה הסטנדרטית נדרשו להיות מאוד ספציפיים לדאטהסט, למשל ״תמונה בשחור לבן של {}״ בעוד ש-CuPL Single נדרשה רק מילה אחת נוספת לתיאור - סוג הקטגוריה, כפי שצויין קודם.

בשיטת CuPL full אמנם נדרש מספר גדול יותר של פרומפטים ידניים מאשר ב-single אבל עדיין הרבה פחות מאשר השיטה הסטנדרטית (כדוגמה - 5 לעומת 80 ב-ImageNet). בנוסף מוצג שיפור משמעותי של מעל לאחוז בחלק מהדאטהסטים.

מעבר לתוצאות הסיווג הכלליות נבחנו גם השפעות של נושאים נוספים על הביצועים:

גודל המודל - לשם בחינת נושא ה-scaling, בוצעה השוואה בין השיטה הסטנדרטית לשיטת ה-CuPL full עם ImageNet עבור גדלי מודל שונים - גם של מודל השפה (ימין) וגם של מודל הסיווג (שמאל). במקרה של מודל השפה הגרף של השיטה הסטנדרטית קבוע מאחר והיא לא עושה בו שימוש. המסקנה היתה שמודלים גדולים יותר יובילו לאחוזי דיוק גבוהים יותר.

כמות הפרומפטים - בוצעה השוואת ביצועים על גדלים שונים של כמות הפרומפטים (נבדקו בנפרד גם פרומפטי ה-LLM וגם פרומפטי התמונה). במקרה של פרומפטי ה-LLM ניתן לראות שיש שיפור כבר החל מפרומפט יחיד. במקרה של פרומפטי התמונה, עבור מספר קבוע של חמישה פרומפטי LLM נוסו גדלים שונים. נמצא כי החל ממספר של 25 פרומפטי תמונה, הגישה החדשה מתעלה על הביצועים של ה-baseline (אשר כאמור, משתמש בכ-80 פרומפטים ידניים ל-ImgeNet).

מגוון פרומפטים - על ידי עדכון ה-temperture של GPT-3 אשר איפשר שילוב של מידת רנדומליות ביצירת הפרומפטים, נבחנה ההשפעה של המגוון. גם כאן הוצג שיפור ככל שהפרומפטים היו מגוונים עבור ImageNet.

השוואה בין שימוש בפרומפטים לשימוש בהגדרות קטגוריית תמונה - הגישה המוצעת נבחנה גם ביחס לשימוש בטקסט תיאורי שמקורו בהגדרת הקטגוריה. מאחר ו-ImageNet מבוסס על WordNet ניתן לחלץ לכל קטגוריה את ההגדרה שלה ועליה לבצע עיבוד לפורמט מתאים. הביצועים שנרשמו פחות טובים מהגישה המוצעת:

נ.ב.

המאמר מציג שיטה אלגנטית וטבעית ליצירה של פרומפטים למשימת סיווג VOC. השיטה מנצלת מודל שפה מאומן ליצירה של מגוון פרומפטים ממספר מועט של פרומפטים כלליים שהונדסו באופן ידני.

שיתוף פעולה: הפוסט נכתב על ידי מיכאל (מייק) ארליכסון, PhD, Michael Erlihson וליאור כהן, Lior Cohen.