# Review 125: Convolutions Die Hard: Open-Vocabulary Segmentation with Single Frozen Convolutional CLIP, 15.08.23 https://huggingface.co/papers/2308.02487

**Paper: https://arxiv.org/abs/2308.02487v2**

בד״כ סגמנטציה בתמונות מתבצעת ב-2 שלבים. קודם מחשבים מסכות עבור כל האובייקטים בתמונה ובשלב השני מזהים סוגי האובייקטים. בזמן האחרון יש שימוש רב במודלים מאומנים(כמו CLIP) להפקה של ייצוג התמונה; בשלב 1 מזינים את התמונה למודל המאומן ובשלב 2 מזינים אותה יחד עם המסכות.   
  
היום ב #shorthebrewpapereviews סוקרים מאמר המבצע זאת בשלב אחד. למה זה טוב בעצם? כי במקרה הזה צריך להזין את התמונה ל-CLIP רק פעם אחת שזה מקצר משמעותית את זמן ההסקה והאימון כי CLIP זה מודל גדול וכבד. איך הם עשו זאת?   
  
קודם כל מעבירים את התמונה דרך CLIP מוקפא ואז מזינים את הייצוג המופק איתו ל-Pixel Decoder יחד עם ״שאילתות האובייקטים״ (סוג של פרומפט לחיפוש האובייקט) לחיזוי המסכות. במהלך האימון מבצעים התאמה בין המסכות ground-truth לבין המסכות שהוצאנו באמצעות אלגוריתם התאמה הונגרי (מזווגים מסכות הדומות ביותר). לאחר שבנינו את המסכות אנו צריכים לזהות את התוכן בתוך המסכות שמצאנו.   
  
עבור סגמנטציה עם מילון סגור (הקטגוריות ידועות) מצליבים את הייצוג (שיכון) של הקטגוריה במסכה (המופק באמצעות הפעלת רשת mask pooling) על הפלט של pixel decoder) עם ייצוג הטקטס (המופק עם CLIP) של כל קטגוריה במטרה למצוא קטגוריה הטובה ביותר לכל מסכה. זה נעשה באמצעות חישוב דמיון cosine (עם טמפרטורה נלמדת) בין ייצוגים אלו כאשר קטגוריה עם דמיון מקסימלי עם ייצוג נבחרת כקטגוריה של המסכה.   
  
כדי לאפשר אוגמנטציה עם מילון פתוח (עבור קטגוריות שלא אומנו במהלך האימון) המחברים יוצרים ייצוג המסכה (מוסיפים ״נתיב״ למודל המקביל לנתיב המילון הסגור) רק מהשיכון המופק מהזנת התמונה ל-CLIP (דרך mask pooling). ואז ב-inference משתמשים באותה שיטה שהסברנו עבור המילון הסגור. בשלב משלבים את החיזוי עבור המילון הסגור והפתוח דרך geometric ensemble (סוג של מיצוע).