# Review 69: [Short] SIMVLM: SIMPLE VISUAL LANGUAGE MODEL PRE-TRAINING WITH WEAK SUPERVISION

**Paper: https://arxiv.org/abs/2108.10904v3**

בהתחלה רציתי לכתוב סקירה קצרה בסגנון #shortdeepnightlearners אבל למעשה ניתן לתקצר את המאמר הזה בכמה משפטים בודדים.

אתם בטח זוכרים את CLIP ו- DALL-E שהצליחו להפיק ייצוגים חזקים (ומתואמים!!) של דאטה ויזואלי וטקסטואלי באמצעות אימון משותף באמצעות שימוש בגישת הלוס הניגודי. כל זה כמובן נעשה על תמונות עם כותרות מהאינטרנט. מחברי SimVLM בחרו בגישה פשוטה יותר (לדעתי) - הם מאמנים רשת שבהינתן תמונה והתחלת הכותרת שלה, חוזה את החלק השני של הכותרת.

איך עושים זאת? מעבירים פאצ'ים של תמונה דרך האנקודר של הטרנספורמר יחד עם החלק הראשון של הכותרת. צריך לציין שאת הפאצ'ים מעבירים קודם דרך רשת באקבון קונבולוציונית (שלום inductive bias). לאחר מכן חוזים את החלק השני של הכותרת עם הדקודר של הטרנספורמר.

הייצוגים המופקים בשיטה אימון הצליחו להגיע לביצועים טובים במגוון משימות כגון סגמנטציה ומענה על שאלות לתמונה נתונה (open visual question answering) ב- zero-shot. גם לתת כותרת לתמונה הרשת יודעת לתת בצורה די טובה ואך זה לא מפתיע…

מאמר: https://arxiv.org/abs/2108.10904

פרויקט: https://ai.googleblog.com/2021/10/simvlm-simple-visual-language-model-pre.html

רב טוב חברים,