# Review 172: [Short] Localizing and Editing Knowledge in Text-to-Image Generative Models

**Paper: https://arxiv.org/abs/2405.01008v2**

https://huggingface.co/papers/2310.13730

מודלי דיפוזיה ממשיכים לשלוט ב AI גנרטיבי כבר זמן מה ואחד נושאי המחקר החמים ביותר בנושא הזה הוא עריכת תמונות המגונרטות עם מודלים אלו. לאחרונה יצאו לא מעט שיטות שמצליחות למשל להוריד אובייקט מתמונה, להחליף אותו לאובייקט אחר או לשנות את סגנון התמונה. המאמר שנסקור היום מציע שיטה לעריכת תמונות המגונרטות עם מודלי דיפוזיה בצורה מאוד אלגנטית המתבסס על ההבנה של מה שקורה בתוך מודל הדיפוזיה (שזה אנקודר של טקסט ומודל המסיר רעש מתמונה UNet בכל איטרציה).

כלומר בשלב הראשון המאמר מנסה להבין איזה חלק(שכבה) במודל להסרת הרעש אחראי על יצירה של כל אובייקט בתמונה, איזו שכבה אחראית על הסגנון, ואיזו מהשכבות אחראית על צבע. איך עושים זאת? קודם כל מוסיפים את הרעש לטוקן האחרון של האובייקט/סגנון/צבע בתיאור הטקסטואלי. למה אותו דווקא?

המאמר בדק ומצא (על ידי השימוש ב Clip-Score המודד את איכות התמונה המגונרטת והתאמתה לתיאור) שזה מה שמשפיע על הישות שרוצים לערוך (למשל מעלים אובייקט). אז איך עושים עריכה? מכיוון שהשכבה הראשונה אחרי שכבת האמבדינג באנקודר היא קריטית אז מאמנים רק אותה (את חלקה). מכיוון שיש לנו טרנספורמרים כאן אז השכבה מוגדרת על ידי 4 מטריצות: W\_q, W\_k, W\_v ו- W\_out.

שלוש המטריצות הראשונות הן מטריצות ממנגנון ה-attention ומשאירים אותן כמו שהן ומאמנים רק את W\_out (לצורך עריכה) תוך כדי שימוש בשיכונים (embeddings) של האובייקט (או סגנון) הישן והחדש c\_k ו-c\_v בהתאמה. פונקצייה שמאפטמים אותה כדי למצוא את W\_out מרמזת על כך שהמטרה(לא לגמרי הפנמתי מה הרציונל כאן) היא למצוא W\_out חדשה כך שפלט של השכבה הראשונה ״החדשה״ עבור c\_k (הישן) תהיה כמה שיותר קרובה לפלט של השכבה המקורית עם c\_v (החדש) עם רגולריזציה קטנה. והכי כיף שניתן לפתור בעיה זו בצורה סגורה ואין צורך באימון שזה מגניב. לבסוף הם עשו עוד דבר נחמד.

הם מצאו שיש שכבה מסוימת במודל להסרת הרעש שאם מעתיקים את האקטיבציות שלה עבור הקלט הטקסטואלי הלא מורעש האובייקט ״הנערך״ חוזר לתמונה המגונרטת. שימו לב שמכיוון של הארכיטקטורה של המודל מבוססת על ResNet זה התוצאה מההזנה של הקלט המורעש לא זהה לזו של לקלט הלא מורעש. אבל כן מקבלים תמונה דומה עם אותו האובייקט. וכמובן ששכבות שונות אחריות על שינוי צבע, סגנון וכדומה.