# Review 23: Make-A-Video: Text-to-Video Generation without Text-Video Data

**Paper: https://arxiv.org/abs/2311.10982v1**

## Make-A-Video: Text-to-Video Generation without Text-Video Data

## Dreamix: Video Diffusion Models are General Video Editors

פינת הסוקר:

המלצת קריאה ממייק: מומלץ לאלו העוסקים בגנרוט ובעריכה של סרטוני וידאו

בהירות כתיבה: בינונית ל- Make-A-Video

ידע מוקדם:

הבנת יסודות של DDPM (כולל איך מגנרטים וידאו מטקסט)

עקרונות של גנרוט של סרטוני וידאו

יישומים פרקטיים:

גנרוט ועריכה של וידאו מטקסט או מתמונה/ות נתונה/ות

פרטי מאמר:

לינקים למאמר: זמין להורדה1, זמין להורדה 2.

פורסם בתאריך: 29.09.222 ו-02.02.2023 , בארקיב.

הוצגו בכנס: ---.

תחומי מאמר:

מודלים גנרטיביים

מודלי דיפוזיה (DDPM)

גנרוט וידאו מטקסט

כלים מתמטיים, מושגים וסימונים:

מודלי דיפוזיה (Denoising Diffusion Probabilistic Models- DDPM)

טרנספורמרים

CLIP

Imagen

DreamBooth

מבוא:

זוהי הסקירה הראשונה ב-deepnightlearners שבה אסקור יותר ממאמר אחד. Make-A-Video הוא מאמר די עתיק (יצא לפני 9 חודשים) הוא אחד המאמרים הראשונים בנושא המרתק של Text2Video. המאמר השני Dreamix משדרג את היכולות שפותחו ב-Make-A-Video ומציע יכולות חדשות כמו עריכה של וידאו וגנרוט של וידאו מכמה תמונות.

נתחיל מלסקור את Make-A-Video שמטרתו היא לקחת שיטות ליצירת תמונות מתיאור טקסטואלי (DALLE-2, Imagen, Stable Diffusion וכדומה) ולהתאים אותן ליצירת סרטוני וידאו. למעשה המאמר מקנה למודל Text2Image יכולת ליצור סדרה קוהרנטית של תמונות (פריימים) במילים אחרות סרטון וידאו.

מודלי Text2Image מבוססי דיפוזיה:

נתחיל מהסבר קצר איך עובדות השיטות לגנרוט דאטה מטקסט (Text2Image) כמו אלו שהזכרנו קודם. מודלים אלו מבוססים כולם על מודלי דיפוזיה הסתברותיים (DDPM).

איור 1: אבני בניין של Stable Diffusion

לרוב מודלי Text2Image מורכבים מהחלקים הבאים:

מקודד של טקסט: (Text Encoder): מפיק שיכון (embedding) של תיאור התמונה. בדרך כלל לוקחים מודל שפה ענקי (LLM) כמו T5 או CLIP ומשתמשים בו כמו שהוא או מכיילים במהלך אימון. המאמר משתמש במודל CLIP מאומן לחישוב ייצוג(=שיכון) של הטקסט ושל התמונה.

מודל הממפה שיכון של טקסט לשיכון של התמונה, Prior. מודל זה הינו אופציונלי ולמשל Imagen לא משתמש בו. Prior בדרך כלל מבוסס על מודלי דיפוזיה (כמו ב-DallE-2 ו-Stable Diffusion) עם ארכיטקטורת U-Net. דרך אגב DallE-2 בדקו ארכיטקטורה אוטורגרסיבית אך לטענת המחברים זה עבד פחות טוב.

מפענח (decoder): בונה את התמונה מהשיכון שלה. מודל זה יכול להיות מבוסס על מודל דיפוזיה (כמו ב-DallE-2 ו-Imagen) או סתם רשת נוירונים כמו ב-Stable Diffusion. שימו לב שהדקודר מבוסס מהווה מה שנקרא cascaded model, שמורכב מכמה מודלי דיפוזיה המבצעים דגימת יתר (upsampling) כל הדרך לרזולוציה הנדרשת.

איור 2: Imagen - מודל דיפוזיה cascaded

תקציר מאמר Make-A-Video:

כאמור Make-A-Video הינו שכלול של מודלי Text2Image המאפשר לגנרט וידאו מהתיאור הטקסטואלי. הבעיה המהותית ביותר באימון מודל כזה היא היעדר דאטהסטים גדולים וזמינים של סרטוני וידאו באיכות גבוהה בעלי תיאור טקסטואלי. המחברים ״עוקפים״ בעיה זו ומאמנים את המודל שלהם קודם כל על הדאטהסט המכילים תמונות עם כותרת ולאחר מכן מכיילים אותו עם סרטוני וידאו לא מתויגים (ללא תאור).

ארכיטקטורת Make-A-Video די מזכירה את מה שתואר בסעיף הקודם. השוני המהותי היא החלפה של קונבולוציות דו-מימדיות (2d) רגילות שהיו ב- UNet (עם attention) שכיכבו ב- Stable Diffusion (ראה איור 3), בקונוולוציות 3d פריקות (separable). במאמר קונבולוציות אלו נקראות spatial-temporal.

איור 3: UNet עם שכבות ה-attention

המימד הנוסף שהתווסף שכבות הקונבולוציה הוא באופן לא מפתיע מימד הזמן. שכבות אלו מאומנות (מכוילות) לשמירת הקוהרנטיות והשטף הרציף של הפריימים (ראה איור 4).

גנרוט וידאו עם Make-A-Video מורכב מכמה שלבים:

יצירת 16 פריימים בגודל 64x64 מהטקסט.

ביצועי דגימת יתר למספר פריימים לשניה (frame rate - fps) גבוה יותר תוך שמירה על הרזולוציה של כל פריים.

שלב SRl: מבצעים דגימת יתר במרחב הפיקסלים ו/או במרחב הזמן (מעלים מספר פריימים לשניה).

בשלב SRh מגדילים רק את רזולוציית הפריימים תוך שמירה על אותו fps (אחרת זה יוצא כבד מדי מבחינה חישובית לאמן את זה). בסוף התהליך מקבלים וידאו ברזולוציה 768x768 בכמה fps אפשריים (המחברים מבצעים downsampling ל-512x512 בטענה שזה מוריד רעשים של תדרים גבוהים שיצאו לא טוב ברזולוציה 768x768).

איור 4: דיאגרמה כללית של Make-A-Video

הערה: איור 4 הינו לגמרי ברור אך ההסבר המעמיק מגיע כבר בפרק הבא.

איך מאמנים Make-A-Video:

האימון של המודל מכיל 3 שלבים עיקריים:

אימון של רשת Prior המקבלת ייצוג (שיכון) של טקסט לשיכון של תמונה, כלומר ממפה שיכון של טקסט לשיכון של תמונה. מודל זה מאומן על דאטהסט של תמונות עם כותרת כאשר השיכון של טקסט והשיכון של תמונה מחושבים עם המודלים המאומנים של CLIP (לא ברור אם מודלים CLIP מאומנים או מוקפאים בשלב הזה). הפריור הוא מודל דיפוזיה מותנה (conditional diffusion model) כאשר מקבל שני פלטים: השיכון של CLIP עבור הטקסט וגם השיכונים של הטוקנים (BPE). לאחר מכן ה-Prior מוקפא ולא מכויל יותר.

מאמנים את הדקודר (היוצר תמונה ברזולוציה הנמוכה ביותר 64x64) ושני המודלים של upsampling על דאטהסט המכיל תמונות בלבד. נציין כי הקלט לדקודר מחושב עם CLIP (מוקפא למרות שהמאמר לא מציין זאת במפורש). בשלבים אלו מודל מאומן ללא שכבות קונבולוציה עתית (temporal convolution) אלא כמו מודל רגיל לגנרוט של תמונות מטקסט. כל המודלים אלו הינם מודלי דיפוזיה מותנים. הדקודר מקבל את הפלט של CLIP (השיכון של התמונה) כקלט והקלטים לשני מודלים של super-resolution הם הפלטים של השלב הקודם.

מוסיפים למודל המאומן בשלב הקודם קונבולוציות עתיות בדקודר ושתי רשתות upsampling ומכיילים אותם עם קטעי וידאו בלבד (ללא כותרת). שכבות הקונבולוציה מתחילות עם המשקלים מהשלב הקודם. בשלב הראשון המודל מייצר 16 תמונות לכותרת נתונה ולאחר מכן מבצעים upsampling (כלומר אינטרפולציה ״רציף״) לתמונות אלו כדי להפוך אותם לוידאו קוהרנטי. בהתחלה משכפלים את הפריימים לקצב פריימים הרצוי ומאתחילים את הקונבולוציות העיתית עם וקטורי one-hot (שקול להפעלת טרנספורמציית זהות/לא לעשות דבר בציר הזמן). כאן האימון מתבצע באמצעות מיסוך של פריימים מסוימים בוידאו כאשר המודל מתבקש לנחש את הפריימים הממוסכים.

תקציר מאמר Dreamix:

המאמר מציע מודל המסוגל לבצע מספר פעולות עם וידאו:

עריכת וידאו בהתאם לתיאור טקסטואלי:

יצירת וידאו מתמונה בודדת ותיאור טקסטואלי (המתארת פעולה בוידאו):

יצירת וידאו מכמה (מספר קטן) של תמונות ותיאור טקסטואלי:

הכותבים משתמשים במודל שהוצע במאמר Imagen-Video כמודל בסיס ומכיילים אותו כדי לאפשר ביצוע של הפעולות הנ״ל. עקרונות המודל של Imagen-Video די דומים ל-Make-A-Video שאותו תיארנו בפרקים הקודמים. שני מודלי אלו מבוססים על אותם העקרונות אך יש להם שלושה הבדלים עיקריים:

Imagen-Video משתמש במודל T5-XXL לבניית שיכון של פרומפט (Make-A-Video משתמש ב- CLIP)

Imagen-Video לא משתמש ברשת Prior ההופכת שיכון של פרומפט לשיכון של תמונה (וידאו למעשה)

Imagen-Video משתמש ב-3 מודלים של upsampling מרחבי (רזולוציה) ועתי (מספר פריימים בשניה) ולא בשניים כמו Make-A-Video.

Dreamix לוקח את המודל של Imagen-Video ומשתמש בו כדי לערוך וידאו וליצור וידאו מתמונה בודדת ותיאור. איך זה נעשה? למשל למשימת עריכת וידאו, קודם כל מוסיפים רעש לווידאו, המתאים לאיטרציה t (ראה הערה מתחת לפסקה זו) מסוימת של התהליך הקדמי של DDPM ומורידים לו רזולוציה מרחבית וגם קצב פריימים בשניה. לאחר מכן משחזרים את הסרטון המשובש הזה לרזולוציה רצויה (המרחבית ו-frame rate) כאשר אחד הקלטים הוא הפרומפט. למעשה מודל Imagen-Video משמש להסרת הרעש ולדגימת יתר של הוידאו המורעש ו- downsampling עד ליצירה קטע הוידאו מקורי. כדי ליצור וידאו מתמונה בודדת ותיאור פעולה, הכותבים משכפלים את התמונה כמה פעמים. לאחר מכן ״מכניסים לתמונות אלו קצת דינמיקה״ על ידי הפעלה של טרנספורמציות המדמות תנועה רציפה (שינוי מיקום, כיוון וכדומה) של המצלמה. לאחר מכן מוסיפים רעש ומעבירים דרך Imagen-Video ליצירה של וידאו נקי ברזולוציה הרצויה.

הערה: למעשה t זה לא מספר האיטרציה אלא מספר בין 0 ל-1 כאשר 0 מסמן דוגמא נקיה ללא תוספת רעש ו-מתאים לרעש גאוסי טהור - האיטרציה האחרונה של התהליך הקדמי).

הגישה הנאיבית הזו, שלא דורשת כיול של Imagen-Video לא הצליחה ליצור וידאו באיכות מספיק טובה עבור משימת עריכת הוידאו. הסיבה לכך לדעת המחברים היא אי יכולת לשמור על התכונות של הוידאו המקורי. כדי להתגבר על בעיה זו, המחברים השתמשו בשיטה שהוצעה ב-DreamBooth (חוקרים ישראלים נטלו חלק במחקר הזה). הכותבים של DreamBooth פיתחו שיטה לכייל מודל דיפוזיה Text2Image מאומן כך שיוכל לגנרט תמונות עם אובייקט נתון (נגיד כלב או אגרטל, ראה איור למטה) - תהליך זה נקרא personalization.

כיול זה מתבצע באופן הבא:

בוחרים ״שם״ לאובייקט באמצעות קידודו עם טוקנים נדירים, כלומר טוקנים בעלי שכיחות נמוכה (ראה הערה)

מכיילים מודל עם תמונות של האובייקט כאשר הפרומפט מכיל את ״השם״ הנבחר של האובייקט (in a doghouse "#$" ).

הערה: בחירה של טוקנים רגילים לקידוד האובייקט ולא קומבינציות נדירות של טוקנים רגילים (כמו xzy30sbs4a) מנומקת במאמר (DreamBooth) באופן הבא. טוקנים רגילים משוכנים (embedded) בנפרד ע"י הטוקנייזר ולכל אחד מהם עשוי להיות פריור חזק. וכיול המודל עלול לשנות אותם וזה עלול לעוות את המבנה של המרחב הלטנטי של הטקסט והיישור (alignment) שלו עם המרחב הלטנטי של התמונות.

המחברים של Dreamix משתמשים בגישה זו ומכיילים את Make-A-Video עם הוידאו המקורי. כאן הכיול משלב את שני היעדים הבאים:

״מלמדים״ את Make-A-Video לשחזר את הוידאו עצמו מהגרסאות המורעשות שלו.

לכל פריים בודד ״מלמדים״ את המודל לשחזר כל פריים נקי מהפריימים המורעשים אחד אחד.

שני כיולים אלו מתבצעים פעם אחת לכל וידאו הנערך ולאחר מכן ניתן לערוך אותו כרצוננו. למעשה זה שלב מקדים לפני עריכה של וידאו נתון.

הערה: לא ברור איך בוחרים את הטוקנים הנדירים לשלבים האלו (האם פריימים של הוידאו וידאו עצמו מקודד עם אותם טוקנים נדירים).

דגימה מ-Dreamix:

כמו ברוב המוחלט של מודלי הדיפוזיה הדגימה בהם מתבצעת באמצעות שיטת DDIM. שיטה זו מאפשרת תהליך מהיר יותר (פחות איטרציות) של הסרת הרעש (denoising= backward process). האצה זו מתאפשרת כאשר משנים את התהליך הקדמי להיות לא מרקובי (xt תלוי לא רק ב xt-1 אלא גם ב-x0). המאמר לדעתי מקבל הרבה פחות תשומת לב ממה שמגיע לו ואני מתכנן לסקור אותו באחת הסקירות הבאות.

נ.ב.

בסקירה הארוכה זו סקרנו שני מאמרים המציעים מודלים text2video. המאמר הראשון מציע שיטה לגנרט וידאו מטקסט, כאשר המאמר השני משכלל יכולות אלו ובונה מודל המסוגל לערוך וידאו בהתאם לתיאור הטקסטואלי וגם ליצור וידאו מסדרה של תמונות.