# Review 153: [Short] FOLEYGEN: VISUALLY-GUIDED AUDIO GENERATION, 24.09.2023

**Paper: https://arxiv.org/abs/2309.10537v1**

https://huggingface.co/papers/2309.10537

בחרתי לסקור את המאמר הזה כי למדתי ממנו שיש משימה שלא הכרתי בראייה ממוחשבת והיא הפקה סדרת אודיו מסרטון וידאו נתון. מתברר שמדובר במשימה לא טריוויאלית והמאמר מציע גישה אלגנטית ודי פשוטה לבעייה זו.   
  
אז היום ב- #shorthebrewpapereviews סוקרים מאמר המציע שיטה להפקה אודיו מוידאו. בהינתן דאטהסט המכיל זוגות של סרטוני וידאו או אודיו המתאים המחברים משתמשים באנקודר מאומן של אודיו EnCodec שהופך את האודיו לייצוגו הלטנטי. מה זה ייצוג לטנטי של אות אודיו?   
  
למעשה זו סדרה של וקטורים שכל אחד מהם הוא השיכון (embedding) של מקטע (בזמן) של האות. בנוסף יש ל-EnCodec דקודר שמשחזר את האות מהייצוג הלטנטי שלו. המאמר גם משתמש במודלים שמטרתם להפיק ייצוג של וידאו (של כל פריים) כמו CLIP, ImageBind ו- ViT.   
  
אז מה בעצם הארכיטקטורה של FoleyGen ואיך מאמנים את המודל הזה? לכל זוג של וידאו ואודיו מעבירים את האודיו דרך האנדקור של EnCodec ואת הוידאו דרך האנקודר של דאטה ויזואלי (נגיד CLIP). כלומר כאן אודיו וידאו מויצגים באמצעות סדרה של וקטורי הייצוג של ״הטוקנים״ שמרכיבים אותם (פריים לוידאו ומקטע זמן לאודיו).  
  
לאחר מכן מאמנים טרנספורמר (מורכב מדקודר בלבד) שמטרתו לשחזר את ייצוג הטוקן הבא של אות אודיו בהינתן ייצוגי הטוקנים (של אודיו) הקודמים וייצוגי של טוקני הווידאו. הם בחנו כמה אופציות לגבי טוקנים של הווידאו של הטרנספורמר יכול לגשת: כל הטוקנים, רק הטוקנים שבאו לפני הזמן או את טוקני הווידאו הסמוכים בזמן. וזה וזה – פשוט ואלגנטי.