

SDXL VAE för kompression av bilder

DVA305

Information – kunskap – vetenskap – etik

[VT] 2026

- Ayman, Martin och Salah
- Epost och kontaktuppgifter
- Eskilstuna, 2026-01-28

Inledning

Behovet att optimera lagring och läsning av bilder har sedan länge varit en stor och viktig del inom datavetenskapen och således föremål för intensiv forskning.

Utvecklingen av komprimeringsalgoritmer har lett till allt bättre metoder för att effektivt kunna inhämta bildinformationen från lagrings- eller strömningsmediet, och därefter snabbt presentera den för oss människor i visuell form. Samtidigt har det i och med informationssamhällets framväxt uppstått ett allt större behov av att kunna göra filerna så kompakta som möjligt för att vi ska kunna lagra denna växande mängd på våra enheter.

Traditionellt sett har bilder endast behövts redigeras i bildbehandlingsprogram såsom Photoshop eller Gimp. Bilden har behandlats utifrån sin rumsliga struktur med pixeln som viktigaste beståndsdel. Bildredigering har i dessa program framförallt handlat om att skala, rotera eller ändra färgen på enskilda pixlar eller grupper av pixlar.

Huvudsyftet har varit att möjliggöra ett intuitivt sätt att arbeta med bilder för oss människor. I detta sammanhang har populära filformat såsom JPG, PNG och AVIF sakta utformats för att tjäna detta syfte på bästa sätt.

Med framväxten av maskininlärning och djupinlärning har nya sätt att bearbeta bilder introducerats. Istället för att manipulera pixlar direkt i det visuella planet finns nu möjligheter att arbeta med bildens semantiska innehåll i en s.k. latent representation. En ML-modell kan med denna latenta representation som grund generera en, eller flera visuellt olika bilder på skärmen. Eftersom ML-modeller inte behöver tolka bilder utifrån samma strukturella ordning som människor gör, kan bildinformationen omstruktureras på nya sätt som bättre passar modellens syfte. Målet med dessa ML-baserade metoder är inte i första hand att minimera filstorleken, men genom att komprimera bort överflödig information underlättar man beräkningsarbetet och ökar på så sätt även prestandan i modellerna. Ser man dock enbart till filstorlek kontra bildkvalitet är dagens etablerade kompressionsalgoritmer redan mycket effektiva. Till exempel kan dessa med lättet krympa bildens filstorlek upp till 60 gånger jämfört med originalet. Detta med en kvalitet som för blotta ögat är nästintill omöjlig att särskilja från originalbilden. Dessutom kan sådana komprimerade bildfiler öppnas på bara några millisekunder. Detta till trots finns ingen egentlig fördel med att använda dessa bildformat i ML-modeller. I komprimerat tillstånd är bildinformationen kodad på ett sätt som gör det omöjligt att bearbeta bilden vidare utan att först återföra den till sin fulla storlek, varpå alla fördelarna har gått förlorade.

Tack vare att ML-modeller inte behöver ha bilder i ett direkt visuellt format under bearbetningen kan man skapa mer kompakta representationer av bilddata. Mycket av den information som krävs enbart för att återge bilden för ögat kan utelämnas tills modellen är klar och bilden ska visualiseras. Resultatet blir ett kompakt dataformat som ändå behåller det viktigaste för att beskriva vad bilden föreställer. Med ett sådant format kan bilder både ta mindre plats på disken och matas direkt in i ML-modeller utan tunga mellanliggande steg för omkodning och rekonstruktion. Detta kan visa sig särskilt värdefullt när man arbetar med väldigt stora bilder eller stora mängder

bilddata, så som till exempel högupplösta fotografier eller video där traditionell hantering skulle vara ineffektiv.

Frågeställning

Hur påverkas den uppmätta bildkvaliteten och filstorleken när SDXL:s latenta rymd komprimeras med AVIF i jämförelse med PNG samt en okomprimerad referens.

Metod

Vi väljer 20 bilder från [Flickr 8k Dataset](#) (eller alternativt [The Kodak Lossless True Color Image Suite](#)).

Bilderna konverteras med SDXL:s VAE (madebyollin/sdxl-vae-fp16-fix) för att omvandla varje bild till dess latenta rymd, exempelvis $128 \times 128 \times 4 @ f32$ om bilden är $1024 \times 1024 \times \text{RGB}$ (HxWxC).

Istället för flyttal kvantifierar vi värderna som heltal med 8/10/12/16 bitar och sparar som vanligt bildformat (RGBA) inför nästa steg.

Bilddatan komprimeras med PNG respektive AVIF i tre olika kvalitetslägen, låg/medel/hög (vilket ger olika filstorlekar). AVIF väljs då detta format stödjer fyra kanaler (RGBA), till skillnad från JPG (RGB).

Slutligen återskapas varje latent representation från de komprimerade filerna och omvandlas tillbaka med samma VAE till en RGB-bild.

Vi mäter skillnaden uttryckt i PSNR och SSIM mellan den ursprungliga bilden och varje rekonstruktion och jämför dessa med respektive filstorlek.