# VAE kompression med AVIF

## DVA305

## Information – kunskap – vetenskap – etik

## [VT] 2026

Författare:

Ayman, Martin och Salah

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. I och med informationssamhällets framväxt har det uppstått ett allt större behov av att kunna lagra filerna så kompakt som möjligt för att vi ska kunna hantera denna växande mängd av data på våra enheter. Utvecklingen har framgångsrikt lett till nya algoritmer som på effektiva sätt kunnat inhämta och presentera bildinformationen från olika lagrings- eller strömningsmedia.

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 (ML) och djupinlärning (DL) 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 lika (eller olika) bilder på skärmen. Eftersom ML-modeller inte behöver tolka bilder utifrån samma strukturella ordning som människor, 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, utan genom att komprimera bort överflödig information underlättar man beräkningsarbetet och ökar på så sätt även prestandan i modellerna. Ser man enbart till filstorlek kontra bildkvalitet är dagens etablerade kompressionsalgoritmer redan mycket effektiva. Till exempel kan dessa med lätthet 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 som tidigare nämnts inte behöver ha bilder i ett direkt visuellt format under bearbetningen forskas det på att skapa mer kompakta representationer av bilddata som samtidigt behåller semantisk. Mycket av den information som krävs enbart för att återge bilden kan utelämnas tills modellen är klar och den slutliga 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 bearbetning och rekonstruktion. Detta kan visa sig särskilt värdefullt när man arbetar med väldigt stora bilder eller stora mängder

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

## Hypotes

Vi förväntar oss att SDXL latenta representation ska gå att komprimera ytterliggare med vanliga metoder såsom AVIF eftersom dess kanaler inte bara tycks bestå av okorrellerat “brus”. Detta verifieras genom att latenten omformas till ett bildlikt format (RGBA) och öppnas i ett verktyg som GIMP där det då går att urskilja tydliga bildstrukturer. Det tyder på att latenten innehåller spatialt sammanhängande information (kanter, större former, färgövergångar) och därmed uppvisar redundans som nämnda kompressionsalgoritmer är konstruerade att utnyttja.

## Frågeställning

Uppvisar SDXL:s latenta representation tillräckligt med 'bildlika' egenskaper för att vanliga bildkomprimeringsmetoder (AVIF/PNG) ska fungera för att minska lagringsstorleken utan att bildkvaliteten faller samman vid rekonstruktion?

## Metod

Vi väljer 20 bilder från Flickr 8k Datfaset (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 128x128x4@f32 om bilden är 1024x1024xRGB (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.