Ett nytt sätt att mäta fel i maskininlärning

Originaltitel: A Novel Perceptual Metric in Deep Learning

Johanna Engman
Lunds Universitet, NVIDIA
tfy15jen@student.lu.se

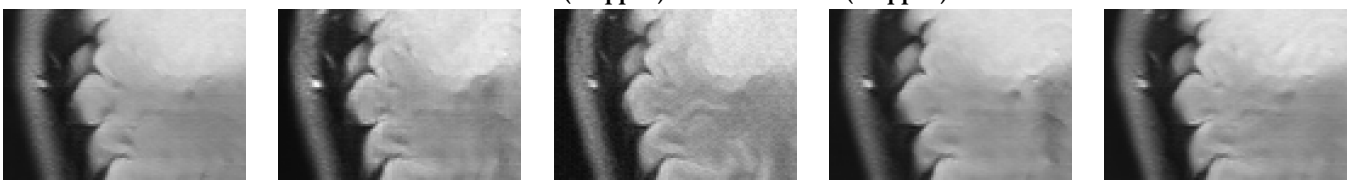
Hanna Nilsson
Lunds Universitet, NVIDIA
tfy15hni@student.lu.se

(a) Originalbild **(b) Brusig input** **(c) Originalbild (croppad)** **(d) Brusig input (croppad)** **(e) l_1**

(f) l_2 **(g) MS-SSIM** **(h) LPIPS+l_2** **(i) FLIP** **(j) Adaptive**

Figure 1: Originalbilder, förvrängda bilder, och olika output från nätverk för färgbilder.

(a) Originalbild **(b) Brusig input** **(c) Originalbild (croppad)** **(d) Brusig input (croppad)** **(e) l_1**



(f) l_2 **(g) MS-SSIM** **(h) LPIPS+l_2** **(i) FLIP** **(j) Adaptive**

Figure 2: Originalbilder, förvrängda bilder, och olika output från nätverk för MRI.

Förmågan att återskapa förvrängda bilder *på ett effektivt sätt har länge varit ett mål inom datorgrafik och bildanalys. I vårt examensarbete, i samarbete med NVIDIA, har vi tittat på ett nytt tillvägagångsätt för detta problem inom maskininlärning och jämfört detta med redan existerande metoder.*

Maskininlärning är ett populärt ämne vars applikationsområden är många. En undergrupp inom maskininlärning är artificiella neurala nätverk (ANN), en typ av algoritm som påminner om uppbyggnaden av den mänskliga hjärnan. Detta eftersom ett ANN består av artificiella neuroner som skickar information mellan varandra. Man ger ett ANN en

stor mängd data (t.ex. bilder på hus, båtar och människor) som den får *träna* på för att kunna lösa olika problem. Två exempel på sådana problem är att ta bort oönskat brus eller återskapa förstörda bilder. Att ta bort brus från bilder är viktigt inom många områden idag, ett exempel är inom spelindustrin där den allt mer avancerade grafiken kräver brusreducering. Att kunna återskapa bilder kan vara väldigt användbart inom sjukvården, t.ex. vid MRI. Detta eftersom färre mätningar krävs vilket leder till lägre kostnader, mindre stress för patienter och möjlighet att utföra MRI inom områden som tidigare har varit för dyra eller långsamma.

Att träna ett ANN innebär att nätverket justerar sig beroende på vilket resultat det får. Man kan jämföra detta mer hur vi människor lär oss i skolan. När en elev gör bra ifrån sig, t.ex. svarar korrekt på frågor, får eleven uppmuntran och positiv feedback från läraren. Om eleven istället svarar fel på ställda frågor får den negativ feedback, och försöker lära sig mer för att nästa gång kunna svara korrekt. Nätverket måste, på samma sätt som en lärare lär ut till ett barn, få feedback på hur bra den hanterade ett problem. Detta görs med en *felfunktion*, som kontinuerligt mäter felet mellan det rätta svaret och nätverkets svar. När modellen har tränat klart vill man oftast återigen mäta skillnaden mellan det rätta svaret och nätverkets svar för att få en slutgiltig utvärdering av nätverkets prestation. Istället för att namnge detta som en felfunktion kallar man det istället för *felmått*. Felfunktionen och felmåttet kan se ut på samma sätt, men kan även vara olika. Skillnaden på dessa kan liknas relationen mellan en lärare och en elev, och ett provtillfälle. En lärares jobb är att kontinuerligt hjälpa eleven i undervisningen, och visa vad som är rätt och fel, samma som en felfunktion. Vid provtillfället är det sedan dags att utvärdera hur bra eleven gör ifrån sig, på samma sätt som ett felmått utvärderar ett nätverks prestation.

Det finns flera flerfunktioner som alternativ idag, vissa mer avancerade än andra. NVIDIA har nyligen utvecklat en ny felfunktion som de kallar FLIP, som försöker efterlikna hur vi människor uppfattar bilder. FLIP är såpass nytt att ingen har använt det som en felfunktion i ett ANN tidigare, vilket motiverade till vårt examensarbete. Vi har utfört en jämförelse på hur bra ett nätverk presterar med olika felfunktioner. Felfunktionerna vi har jämfört FLIP med heter l_1, l_2, MS-SSIM, LPIPS+l_2 och General &Adaptive Robust Loss. I Figur 1 och 2 presenteras våra resultat för brusreduceringen respektive MRI rekonstruktionen.

Arkitekturen på hur vårt nätverk är uppbyggt heter *U-net* som är väl dokumenterat och används flitigt i liknande områden. Vi tränade flera nätverk med samma uppbyggnad men med olika felfunktioner vilket vi refererar till som olika modeller. Input till nätverken var brusiga färgbilder eller förvrängda MRI bilder, beroende vilket problem vi undersökte. Utvärderingen av outputen gjordes i tre steg:

1. Våra egna observationer med ett sammanfattande poängsystem
2. Mätvärden från felmått för modellerna
3. Användarstudier med tillhörande statistik analys

I våra egna observationer visar vi exempel på modellernas output och pekar ut för läsaren där felfunktionerna har lyckats eller misslyckats. Vi sammanfattade sedan våra observationer i ett poängsystem över t.ex. hur mycket brus modellen tog bort och om modellen behöll detaljer i bilden. Vidare mätte vi hur bra modellerna presterade enligt några felmått. Till sist utförde vi två användarstudier för de två nämnda problemen. Användarna fick se två bildpar som flippade mellan originalbilden och output från nätverket, för att tydligt kunna uppfatta var modellen hade lyckats eller misslyckats. Uppgiften var sedan att välja det bildpar som visade högst likhet. Denna jämförelse gjordes för alla modeller i randomiserad ordning, där användaren var omedveten om vilka två modeller som visades i varje jämförelse. Vi utförde sedan en statistik analys på resultaten från användarstudierna för att kunna avgöra om de var signifikanta eller ej.

Vi kom fram till att FLIP var en bra felfunktion för brusreducering, men att den hade svårare med MRI rekonstruktionen. LPIPs+l_2 och MS-SSIM rankades överlag högst i båda problemen, men hade problem med att de skapade falska detaljer, framförallt i MRI rekonstruktionen. Återkommande feedback från användarstudierna var att det var svårt att se en större skillnad mellan bilderna för de olika modellerna och därför välja en vinnare, vilket är värt att ta med när man analyserar resultaten.