PERSONLIGA VIDEOMEDDELANDEN MED HJÄLP AV DEEPFAKE OCH MASKININLÄRNING

POPULÄRVETENSKAPLIG SAMMANFATTNING - Johan Liljegren, Pontus Nordqvist

MED HJÄLP UTAV DE SENASTE RÖNEN INOM MASKININLÄRNING utvecklar vi en produkt, som automatiserat kan producera nästa generation av videomeddelanden. Detta med hjälp av den så kallade "deepfake"-teknologin.

För företag är kundvård en del av att upprätthålla ett gott förhållande till sina klienter. Nu görs det möjligt att få en personlig hälsning från företaget i form av en video som kan säga kundens namn. Produkten som vi föreslår använder sig av en förinspelad video eller en vanlig stillbild, innehållande en person, och en ljudsnutt. Dessa kombineras ihop av vårt nätverk så att personen ser ut att tala ljudinnehållet med synkroniserade läppar.

Deepfake-teknologin, som vi använder, har oftast dålig renomé då det oftast förknippas med dåliga saker så som falska nyheter och porrindustrin. Men faktum är att det kan användas till många bra saker också, som automatiserad videodubbning i filmer och för att förebygga omtagningar vid filminspelning. Men även för översättning av interaktiva läromedel till alla världens språk, för att möjliggöra utbildning för alla barn oavsett bakgrund.

Processen ovan realiseras med maskininlärning i form av ett artificiellt neuralt nätverk. Huvudkomponenten i vårt nätverk är ett ramverk som kallas för "General Adversarial Network" (GAN). Detta ramverk har två stycken beståndsdelar, en generator och en diskriminator. Dessa kan liknas vid en förfalskare, generatorn, och en detektiv, diskriminatorn. Förfalskaren skapar bilder som ska efterlikna en samling riktiga bilder som används som referens. Detektiven har i uppgift att urskilja vilka bilder som är riktiga och vilka som är skapade av förfalskaren. Förfalskaren får hela tiden reda på detektivens beslut och på så sätt kan den anpassa sina förfalskningar så de blir ännu bättre. Detektiven får också reda på om rätt bilden blev korrekt identifierad som falsk eller riktig. Förfalskaren och detektiven arbetar i tandem i en form av katt-och-råtta-lek där båda hela tiden blir bättre på sin uppgift. Detta fortgår tills förfalskarens bilder inte alls går att urskilja från referensbilderna.

I denna tillämpning har även ljud adderats till bilder, det vill säga att generatorn och diskriminatorn ska kunna hantera både ljud och bild samtidigt. Katt-och-råtta-leken brukar benämnas som att nätverket "tränas". Träningen gör att generatorn succesivt blir bättre och detta visas i bild 1.

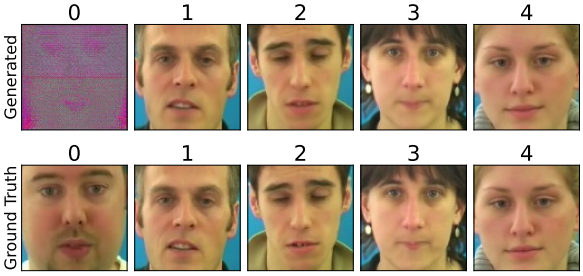

Bild 1: Träningsprocessen visualiserad för varje genomkörning av referensdatan. **Över:** De av generatorn skapade bilderna. **Under:** Referensbild som generatorn försöker efterlikna

Träningsprocessen fortgick inte alltid utan att problem uppstod. Vårt nätverk är stort, komplicerat och känsligt för vilken referensdata vi använder. Ibland uppstod ett problem där generatorn misslyckas med sin uppgift att förbättra sina producerade bilder. Då kan det som kallas för Helvetica-scenariot uppstå, vilket kan leda till oanade konsekvenser. Ett exempel på detta visas i bild 2.

Bild 2: Helvetica-scenariot leder till att generatorn producerar ansikten som kan liknas vid abstrakt konst.

Till slut, efter testning av två olika variationer av vårt ramverk, lyckades vi skapa en modell som gav bra resultat, både i form av genererade bilder och att synkronisera läpparna med ljudet. Modellen fungerar inte helt perfekt och det finns förbättringsområden och förutsättningar för att fortsatt undersöka fältet deepfake för att skapa önskvärda produkter och tjänster.