

Kvalitativ dataurval för AI

Populärvetenskaplig Sammanfattning

För att träna en bildigenkänningsmodell behövs många exempelbilder och en tillhörande klassbeskrivning. Givet att man har ett stort dataset med bilder och klassbeskrivningar (hund, katt, etc.) kan modellen lära sig att gissa klassbeskrivningen på nya bilder den inte sett tidigare. Det är emellertid svårt att få tag på bra klassbeskriven data då lagar som GDPR förhindrar att bilder på människor sparas, då det finns få publika dataset samt ägarskap över bilder kan komma i vägen. Det är även en dyr och långsam process att låta människor skapa klassbeskrivningar på ny data. Ett sätt att lösa detta problem är att använda sig av syntetiska bilder, dessa bilder är datogenererade och associeras med en klassbeskrivning vid produktion. Hur vet vi vad för bilder som ska genereras? Vilka bilder kan snabbt förbättra modellen? Hur många bilder behövs? För att svara på dessa frågor behövs ett sätt att betygsätta datan. I rapporten bygger vi upp en process för detta och motsätter flera metoder för att värdesätta data mot varandra samt utvärderar metoderna på en serie dataset.

I denna rapport diskuteras hur man på ett smart sätt kan utvärdera kvaliteten hos ett dataset som sedan används för att träna en maskininlärningsmodell. För detta projekt har vi arbetat med två olika modeller, en för bildigenkänning och en för objektigenkänning. Vår process för utvärdering av metoder, och följaktligen utvärdering av datan, lyder som följande. Vi har en på förhand tränad modell och en mängd nya bilder. Modellen får då gissa vad dessa nya bilder föreställer, denna gissning används sedan av metoderna för värdesättning av bilderna. Bilderna ordnas efter betyg och en delmängd väljs ut för att läggas till i träningsmängden. Denna träningsmängd används för att på nytt träna modellen och utvärdera dess nya prestanda. Vi repeterar detta tills dess att träningsmängden har alla bilder som vi startade med.

Ovan beskrivna process kallas för aktiv inläring. Denna process visade sig ha många faktorer eller parametrar som påverkade våra resultat avsevärt. För att se till så att resultaten och metoderna för värdesättning av datan skulle vara jämförbara mot varandra kom vi fram till en grundinställning. Eftersom vi ordnade bilderna efter deras betyg, valdes de "svåra" bilderna ut först vilket gjorde det svårare för modellen att lära sig. Därmed slumpas träningsmängden innan modellen tränar. De olika metoderna för värdesättning väljer ut olika "svåra" bilder och modellen kan därmed vara i behov av att träna olika länge, genom att förtidstoppa träningen när modellen slutar se förbättring blir detta möjligt. Vi kom även fram till att det inte nödvändigtvis alltid är fördelaktigt att välja ut de svåraste bilderna. Detta eftersom det förekommer att bilder har fel klassbeskrivning och därmed inte kommer kunna tillföra något till modellen, de är egentligen av låg kvalitet.

Som ett resultat av att förtidstoppa träningen när modellen slutat bli bättre, blev ibland modellen stoppad för tidigt, men bara i början då den hade valt ut lite data. Det visade sig att datan som väljs ut måste vara representativ för problemställningen. Att bara välja ut bilder av bilar kommer

inte att generalisera bra om du vill upptäcka både bilar och människor till exempel. Detta är inte ett problem med många bilder, men vi fann att resulterande figurer ofta blev bättre och stabilare om man introducerade en slumpvis vald bas av bilder. Dessa blir sedan tillförda till träningsmängden i början.

För betygsättning av våra bilder använder vi oss huvudsakligen av två olika former av värdesättning metoder, en tilltro- och en förlustfunktion. Tilltrofunktionen försöker avgöra hur säker eller osäker modellen är på vilken klass en bild borde tillhöra. Medan förlustfunktionen beräknar skillnaden mellan den gissade och den faktiska klassen. Både tilltro- och förlustfunktionen presterade bättre än slumpvis valda bilder, vilket tyder på att ett smart dataurval kan leda till en bättre modell trots mindre data. Tilltrofunktionen kan värdesätta bilder som ännu inte har tilldelats klasser, vilket gör dem väldigt användbara för data prioritering inför databeskrivningsprocessen. Förlustfunktionen presterade bättre än tilltrofunktionen, men kräver att man har klassbeskriven data. Vi såg lovande resultat med dessa metoder på både bild- samt objektigenkänningsproblemet, även om resultaten av objektigenkänningsmodellen visade sig vara svårtolkade.

En iakttagelse var att dataurval med förlustfunktionen ibland presterade bättre trots att modellen inte sett alla träningsbilder. Detta ifrågasätter alltså tanken om att mer data alltid resulterar i bättre resultat. Vi drar slutsatsen att med hjälp av värdesättningsfunktioner och aktiv inläring, kan man välja ut en delmängd av bilder varpå modellen efter träning presterar bättre än om den tränat på alla bilder. Vi presenterar detta som en ny metod att förbättra sin modell innan man börjar använda den i produktion, likt liknande redan existerande metoder som modell reducering och modell komprimering. Vi presenterar även en aktiv inläring-konfigurering för en jämförbar utvärdering av värdesättningsfunktioner, med rimliga och testade konfigureringsparametrar.