Integration av bild- och ordinbäddningar för beskrivande bildlikhet

David Gustafsson och Tobias Lindberg

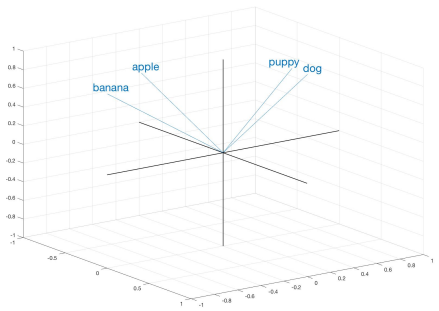
Juni 2017

Bakgrund

Många människor har idag en privat digital fotosamling, i molnet eller på en dator. Sådana samlingar är ofta bara kronologiskt sorterade, även om mer intelligenta lösningar som t.ex. ansiktsigenkänning allt oftare används för att skapa smarta strukturer. Ett annat sätt att göra en fotosamling mer dynamisk och intressant skulle kunna vara genom att föreslå semantiskt relaterade foton till det foto användaren tittar på för tillfället. Dessutom, om den semanstiska relationen kan beskrivas i ord skulle det göra systemet mer transparent och skapa ytterligare värde för användaren. I detta examensarbete utforskas möjligheterna att för det ovan nämnda ändamålet integrera bild- och ordinbäddningar i en gemensam vektorrymd.

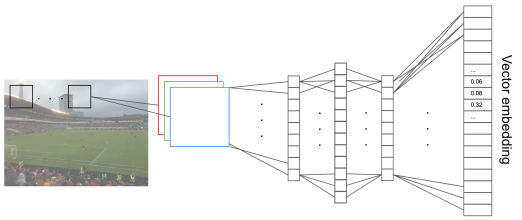
Metod

När datorer hanterar mänskligt språk används numera ofta tekniker som bygger på vektorrepresentationer av ord. Det innebär att varje unikt ord i ett språk är kopplat till en vektor i en t.ex. 300-dimensionell rymd. En fördel med detta är att vektorerna för två ord som har liknande betydelse kommer att ligga nära varandra i rymden, och alltså kan datorn på så vis skapa en uppfattning om ords betydelser snarare än att behandla dem som helt oberoende entiteter. En illustration av detta kan ses i figur 1.



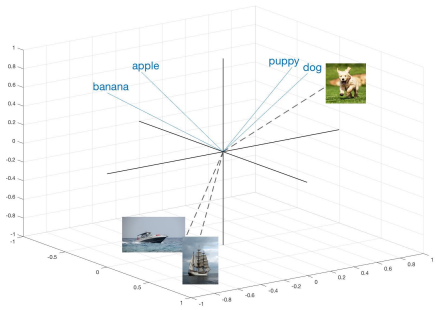
Figur 1: Tredimensionell vektorrymd för ord.

Så kallade Convolutional Neural Networks (CNN) är en teknik som används inom bildseende och som hämtat inspiration från hur den mänskliga hjärnan processar synintryck. Ett CNN tar pixeldatan i en bild som input och kan genom att utföra matematiska operationer i flera lager tränas till t.ex. ansikts- eller objektigenkänning. Värdena mellan dessa lager kan extraheras i form av vektorer och det har visats att sådana vektorer med framgång kan användas för bildhämtning.



Figur 2: Ett CNN som genererar en vektorinbäddning för en bild.

Genom att summera ordvektorerna för en mänskligt komponerad bildtext och träna ett CNN till att förutspå den resulterande vektorn för motsvarande bild kan vektorrespresentationer för bilder skapas i samma 300-dimensionella rymd som vektorrepresentationerna för ord redan existerar. Detta illustreras i figur 2 och 3.



Figur 3: Tredimensionell integrerad vektorrymd för ord och bilder.

Resultat

I examensarbetet visas att den beskrivna metoden lämpar sig väl för de avsedda ändamålen. Genom att ranka bilder från olika grupper baserat på hur semantiskt relaterade de är till en given bild från en av dessa grupper kan rankingkvaliteten mätas och olika modeller jämföras. En grundlig undersökning visar bl.a. att modeller där värden hämtas ur senare lager i det neurala nätverket ger bättre resultat.