Een Steen van Rosetta voor het geautomatiseerd herkennen van digitaal beeldmateriaal

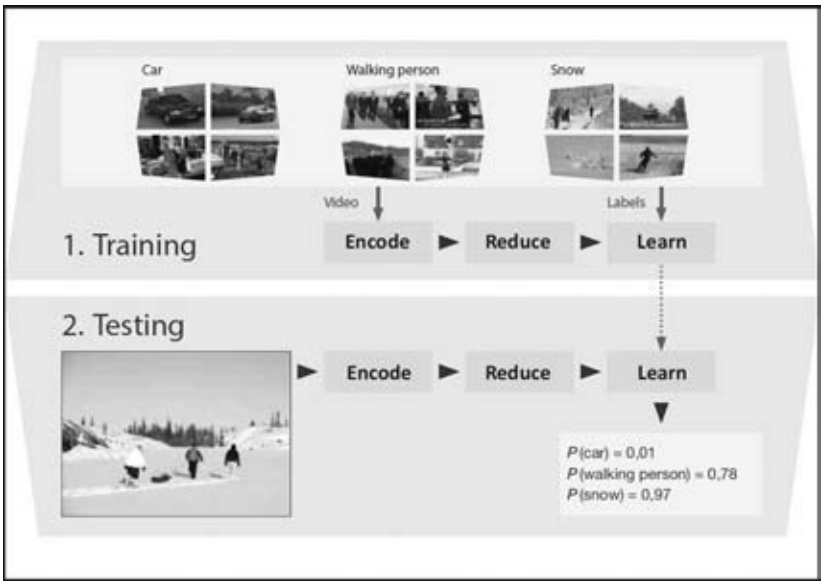
CEES SNOEK

Een klassiek probleem rond het begrijpen van afbeeldingen is de ontcijfering van de Egyptische hiërogliefen. Het gebruik van hiërogliefen als het voornaamste alfabet stopte toen rond 400 na Christus de voorkeur werd gegeven aan het Demotische schrift en het oude Griekse alfabet. Al snel ging de kennis van hiërogliefen volledig verloren. In de eeuwen die volgden probeerden vele wetenschappers de hiërogliefen[1] te ontcijferen, maar het duurde nog tot 1799 voordat echte vooruitgang kon worden geboekt. In dat jaar ontdekten wetenschappers die door Napoleon naar Egypte werden gestuurd een inscriptie in een steen die een vertaling bood van de hiërogliefen in zowel het Demotische als het oude Griekse schrift. Uiteindelijk bleek deze Steen van Rosetta de sleutel waarmee Jean-François Champollion in 1822 de hiërogliefen wist te ontcijferen. Het verkrijgen van inzicht in afbeeldingen in de moderne, DIGITALE tijd heeft veel overeenkomsten met het ontcijferen van hiërogliefen. In dit artikel bespreken we recente onderzoeksinspanningen aan de Universiteit van Amsterdam die tot doel hebben een beeld te vertalen naar de meest beschrijvende samenvatting op concept- en zinsniveau.

1 Afbeeldingen begrijpen door concepten te herkennen

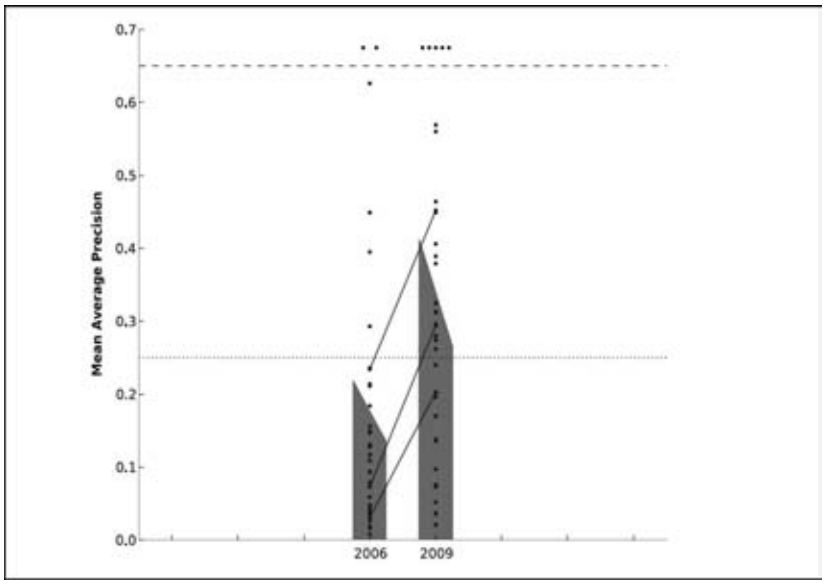
Voor mensen is het begrijpen en interpreteren van een visueel signaal dat de hersenen binnenkomt een ongelooflijk complexe taak. Ongeveer de helft van de hersenen is betrokken bij het toekennen van een betekenis aan een beeldsignaal, om te beginnen met de indeling van alle visuele concepten in het plaatje. Dankzij een aantal doorbraken op het gebied

van computervisie en machine learning, is het categoriseren van de beelden op het conceptniveau is ook binnen handbereik gekomen voor machines. Maar, net zoals bij het ontcijferen van hiërogliefen, komt de grootste bijdrage voor het machine understanding van beelden uit de beschikbaarheid van schriftelijke vertalingen van beeld in de vorm van beeldlabels. De standaard aanpak van machine understanding van beelden begint met een bepaald visueel concept bijvoorbeeld een boot. De set van gelabelde foto's wordt verdeeld in een trainingsset en een testset. De trainingsset wordt gebruikt voor de optimalisatie van het algoritme en voor het aanleren van een zogenaamd statistisch model dat de visuele weergave van het betreffende concept vastlegt in een wiskundige formulering. De testset wordt gebruikt om de mate waarin het model beelden herkent te evalueren door zijn voorspellingen te vergelijken met de oorspronkelijke afbeelding. De tweede stap bouwt een model van een concept. Hiervoor wordt elk beeld geanalyseerd door het onderwerp te extraheren tot een miljoen visuele kenmerken. Deze kenmerken zijn onveranderlijke identificerende elementen die toevallige verschillen in de opname, veroorzaakt door andere belichting, gezichtspunt, of schaal, teniet doen. De derde stap is het projecteren van de identificerende elementen per pixel op een van 4000 woorden. Dit zijn geen echte woorden, maar eerder samenvattingen van een gedeelte van het beeld dat één enkel detail beschrijft: een hoek, textuur, of punt. In de vierde stap zet een machine learning-algoritme de visuele woorden om in de waarschijnlijkheid dat een begrip aanwezig is in een afbeelding. Deze waarschijnlijkheden worden gebruikt om alle beschikbare afbeeldingen te rangschikken naar de aanwezigheid van het begrip. De werkwijze om beelden te begrijpen door hun meest beschrijvende begrippen te herkennen is samengevat in figuur 1. Cruciale stimulansen voor vooruitgang in het geautomatiseerd herkennen van afbeeldingen zijn internationale zoekmachine benchmarks zoals de *TRECVID* (TREC Video Retrieval) benchmark, welke wordt georganiseerd door het *National Institute of Standards and Technology*[2]



Figuur 1. Algemeen schema voor het waarnemen van visuele concepten in afbeeldingen

TRECVID is erop gericht de vooruitgang te bevorderen van het zogenaamde content based retrieval van digitale video via open metrics-based evaluatie. Met de steun van 50 teams uit de academische wereld en het bedrijfsleven, met inbegrip van de Universiteit van Oxford, Tsinghua Universiteit en IBM Research, is het in de praktijk de standaard geworden voor de evaluatie van video retrieval onderzoek. Benchmarks open karakter zorgt voor de snelle convergentie van effectieve benaderingen van beeldbegrip. Onlangs beoordeelden wij de vooruitgang in beeldbegrip door middel van de herkenning van de meest beschrijvende concepten door een state-of-the-art zoekmachine uit 2006 te vergelijken met een uit 2009. We bekeken zowel een situatie waarin de set trainingsgegevens visueel vergelijkbaar was met de testset, dat wil zeggen dat beiden video van hetzelfde genre bevatten, en een situatie waarin de set trainingsgegevens visueel verschilde van de gegevens die werden gebruikt voor testdoeleinden, dat wil zeggen dat zij video's bevatten uit verschillende genres. Zoals figuur 2 laat zien, zijn de prestaties van de zoekmachine verdubbeld in slechts drie jaar.

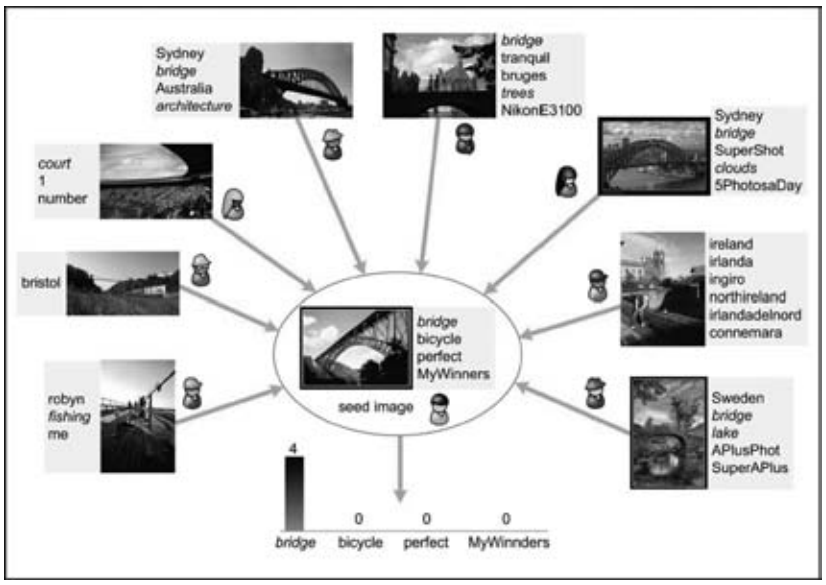


Figuur 2. Vooruitgang prestaties zoekmachines voor visuele concepten, geevalueerd voor 36 detectoren (•)

Voor de herkenning van concepten, namen de vindpercentages af wanneer er gegevens van verschillende oorsprong werden toegepast, maar de resultaten zijn nog steeds verdubbeld in die drie jaar. De vooruitgang mag dan groter zijn dan verwacht, dat betekent echter niet dat het algemene probleem van visueel zoeken is opgelost. Ons experiment gebruikt slechts 36 concepten, terwijl breed beeldbegrip het gebruik van duizenden detectoren zou vereisen om de woordenlijst van een gemiddelde gebruiker te benaderen. Het uitbreiden van de woordenlijst van het begrip naar iets dat de menselijke taal benaderd, vergt een volledige 'Steen van Rosetta' die per begrip honderden vertalingen van afbeeldingen biedt. In de literatuur is het gebruikelijk om voor het verkrijgen van deze vertalingen te vertrouwen op de labeling door experts. Dergelijke labeling is duur en dus beperkt beschikbaar. We hebben onderzocht of het mogelijk is om in plaats van de experts gebruik te maken van niet professionele consumenten die gebruik maken van internetdiensten zoals YouTube en Flickr. Op deze sharing-websites zijn dure kwaliteitslabels op ongekende schaal door gratis labelsvervangen, maar het is bekend dat deze gratis labels overdreven

persoonlijk, ongecontroleerd en dubbelzinnig zijn. Daarom is het een fundamenteel probleem om de relevantie te interpreteren van een door gebruikers gemaakte tag die de visuele inhoud beschrijft. Intuïtief zullen tags waarschijnlijk objectieve aspecten van de inhoud weergeven wanneer verschillende personen visueel vergelijkbaar beelden labelen met behulp van dezelfde tags.

Uitgaand van deze intuïtie, hebben wij een automatisch algoritme voorgesteld, die nauwkeurig en efficiënt label relevantie aanleert door het accumuleren van stemmen van de visuele buren. Het belangrijkste principe achter software voor het automatisch taggen van beeld is erg simpel: Neem een afbeelding, bijvoorbeeld het plaatje in figuur 3, dat is voorzien van de tags, brug, fiets, perfect en MyWinners. Kijk dan naar andere afbeeldingen die een visuele gelijkenis vertonen met de afbeelding. Als de visueel vergelijkbare afbeeldingen in de meeste gevallen ook zijn gecodeerd met dezelfde labels, dan is het vrij waarschijnlijk dat deze code het meest beschrijvende label voor de afbeelding is. In het geval van figuur 3, is dat het label 'brug'. Door mensen aangemaakte tags op het web fungeren als 'Steen van Rosetta' voor beeldherkenning op begrip-niveau.



Figuur 3. Verzamelen gratis gelabelde afbeeldingen van het web met label relevantie van stemmen van visuele buren

2 Beelden begrijpen door zinnen te herkennen

Hoewel de resultaten voor beeldherkenning met behulp van een enkelvoudig concept of begrip indrukwekkend zijn, is er nog weinig bereikt op het gebied van automatische beschrijvingen van gebeurtenissen op het niveau van een hele zin. Dit is niet verrassend, want een gebeurtenis is geen concept. De ideale detector van gebeurtenissen moet een door mensen te begrijpen herkenning opleveren om vast te stellen welke informatie in de video bepalend is voor de relevantie. Het aangeven van 'stoel, stoel, stoel' is niet informatief, het aangeven van 'theater' is dat wel. Niettemin behandelen sommige onderzoekers gebeurtenissen hetzelfde als concepten, waardoor hetzelfde hekenningsproces kan worden toegepast. Maar door het grote aantal sterk gerelateerde eigenschappen en geprojecteerde woorden is het niet eenvoudig af te leiden hoe deze detectoren leiden tot de classificatie van gebeurtenissen. Bovendien worden gebeurtenissen vaak gekenmerkt door overeenkomsten in semantiek, in plaats van in uiterlijke kenmerken. Ons doel is om een informatieve representatie te kunnen vinden die in staat is om gebeurtenissen uit willekeurige video-inhoud te herkennen en uiteindelijk zelfs te beschrijven. We stellen dat een meer semantische representatie noodzakelijk is om die langetermijndoelstelling te halen.

Als eerste stap analyseren we de woordenlijst die mensen gebruiken op webpagina's met video's van gebeurtenissen. Figuur 4 toont een aantal video's en de bijbehorende tekstuele beschrijvingen. Wij verwerken die tekstuele beschrijvingen voor in totaal 13.265 video's. Na een basale tekstuele voorbewerking, zoals het verwijderen van stopwoorden en afgeleide woorden, komen we uit op 5433 verschillende termen. Kijkend naar de menselijke woordenlijst, zien we dat de gebruikte termen kunnen worden ingedeeld in vijf verschillende concepttypes die veel worden gebruikt in multimedia- en computer vision-literatuur: objecten, acties, scènes, visuele kenmerken en niet-visuele begrippen. We schrijven handmatig elke term in de woordenlijst toe aan een van deze vijf types. Na deze oefening zien we dat 44 procent van de termen naar objecten verwijst. Bovendien stellen we vast dat een aanzienlijk aantal objecten is gewijd aan verschillende soorten dieren en mensen, zoals bijvoorbeeld: leeuw en tiener. Ongeveer 21 procent van de termen stellen acties voor, zoals lopen. Ongeveer 10 procent van de concepttypes behandelen scènes, zoals keuken. Visuele kenmerken beslaan ongeveer 13 procent van de termen, bijvoorbeeld: wit, plat en vies. De resterende 12 procent van de

termen beslaan begrippen die niet visueel zijn, zoals: gedicht, probleem en taal. Bovendien zien we dat de woordenlijst zowel specifieke als algemene begrippen bevat. Deze analyse geeft richtlijnen voor het samenstellen van woordenlijsten voor het begrijpen van beelden op zinsniveau.

Figuur 4. Voorbeelden van video's en menselijk toegevoegde tekstuele beschrijvingen

Nadat we de begrippenwoordenlijst die mensen gebruiken om gebeurtenissen te beschrijven hebben gedefinieerd, zijn we klaar voor geautomatiseerde herkenning van beelden op het niveau van een hele zin. Opnieuw is de eerste stap het verzamelen van positieve en negatieve voorbeelden van een bepaalde visuele gebeurtenis, bijvoorbeeld een persoon die een band verwisselt. Vergelijkbaar met de detectie van een concept, worden de gegevens verdeeld in een trainingsset voor ontwikkeling en een testset voor evaluatie. De tweede stap is het bouwen van een model van een gebeurtenis. Hiervoor decoderen we video's door elke twee seconden een frame op uniforme wijze te extraheren. Vervolgens worden alle conceptdetectoren van onze woordenlijst toegepast op de uitgelichte frames. Door de output van elke detector aaneen te schakelen, wordt elk frame vertegenwoordigd door een

begrippenvector. Uiteindelijk worden de REPRESENTATIES (vertegenwoordigingen) van de frames door middeling en normalisatie samengevoegd tot een videoniveau. Boven deze representatie per video van de begrippenwoordenlijst, gebruiken we in de derde stap opnieuw een machine learning algoritme dat de begrippenwoordenlijst omzet in een gebeurtenis score. Deze waarschijnlijkheden worden gebruikt om alle beschikbare videos te rangschikken op basis van de aanwezigheid van een gebeurtenis. Net als bij de herkenning van concepten, zijn de gelabelde voorbeelden van gebeurtenissen van cruciaal belang voor het beschrijven van de afbeelding op niveau van een hele zin. De vraag welke automatische detectors in de woordenlijst voor geautomatiseerde zinsbeschrijving van beelden moet worden opgenomen is nog onderwerp van discussie. Recent onderzoek aan de Universiteit van Amsterdam heeft gezocht naar een constructie voor de geautomatiseerde beschrijving van afbeeldingen op het niveau van een hele zin. Natuurlijk kan men uitgaan van een basiswoordenlijst die zoveel begrippendetectoren bevat als men maar kan bedenken, maar men kan ook te weten zien te komen welke concepten in een woordenlijst de meeste informatie geven over een gebeurtenis. In een recent artikel hebben we het selecteren van een woordenlijst van informatieve concepten uit een grote set van begrippendetectoren voorgesteld als (als) het zoeken naar een zeldzame gebeurtenis. De oplossing die wij bij benadering voorstellen vindt de optimale conceptenwoordenlijst met behulp van een cross-entropie optimalisatie. Onze experimenten tonen aan dat 1) sommige conceptenwoordenlijsten voor specifieke gebeurtenissen informatiever zijn dan andere, 2) de detectie van gebeurtenissen met behulp van een automatisch verkregen informatieve conceptenwoordenlijst robuuster is dan het gebruik van alle beschikbare concepten, en 3) de informatieve conceptenwoordenlijsten nuttig zijn, zonder dat ze daarvoor zijn geprogrammeerd. Zie bijvoorbeeld de informatieve conceptenwoordenlijst voor de zin '*a person landing a fish*' in figuur 5. Het lijkt er dus op dat het voor videodetectie van gebeurtenissen met behulp van conceptenwoordenlijsten loont om informatief te zijn. Er is echter nog veel meer vooruitgang nodig voordat beeldbeschrijvingen op zinsniveau net zo nauwkeurig kunnen worden gegegenereerd als concepten.

Figuur 5. Informatieve conceptenwoordenlijst voor de zin: person landing a fish in

Conclusie

In dit artikel bespreken we recente onderzoeken aan de Universiteit van Amsterdam die tot doel hebben een beeld te vertalen naar de meest beschrijvende samenvatting op begrip- en zinsniveau.[3] Wij zijn van mening dat de vooruitgang in kunstmatige intelligentie, met behulp van computer vision en machine learning, in combinatie met de ruime beschikbaarheid van beschrijvingen op het web, fungeren als 'Rosetta Stone' voor beeldherkenning. Het is onze overtuiging dat de generatie van geautomatiseerde metadata voor beelden binnenkort een sprong voorwaarts maakt van enkele woorden tot volledige zinnen.

Noten

1. J. Ray. The Rosetta Stone and the Rebirth of Ancient Egypt. Harvard University Press, Cambridge, MA,2007.
2. http://trecvid.nist.gov.
3. Vrij toegankelijke publicaties beschikbaar op http://www.ceessnoek.info.