Deep learning for visual understanding

Guo, Y.; Guo Y.

Nederlandse Samenvatting

Al geruime tijd is het doel van 'Computer Vision'-onderzoekers een algoritme te ontwikkelen dat in staat is om automatisch en accuraat visuele informatie te begrijpen. Terwijl het lijkt dat mensen dit zonder enige inspanning kunnen is er tot op de dag van vandaag geen robuuste oplossing beschikbaar vanwege de bekende semantische kloof tussen het object of concept dat gemodelleerd wordt en haar low-level features. Recentelijk, zijn 'Deep Learning'-algoritmen effectief gebleken om deze semantische kloof te dichten voornamelijk als gevolg van de geavanceerde visuele representatie die zij hebben ontwikkeld. Dit heeft geresulteerd in substantiële verbeteringen in verschillende visuele toepassingen zoals beeld-classificatie, object-detectie, beeld-ondertiteling, etc. Het doel van deze thesis is om nieuwe 'Deep Learning'-algoritmen te onderzoeken en te ontwikkelen voor een beter begrip van visuele informatie.

Ten eerste geven we een uitgebreid en diepgaand overzicht van recente ontwikkelingen en vooruitgang op het gebied van 'Deep Learning'. Speciaal gericht op onderzoekers in de gebieden 'Neural Computing', 'Computer Vision' en 'Multimedia' die geïnteresseerd zijn in de state-of-the-art in 'Deep Learning in Computer Vision'. Vervolgens beschrijven we ons onderzoek op het gebied van drie visuele toepassingen: traditionele beeld-classificatie, hiërarchische beeld-classificatie en beeld-beschrijving (-'captioning').

De traditionele beeld-classificatie-taak bestaat uit het classificeren van een beeld in een van te voren gedefinieerde categorie. Dit probleem is binnen de 'Computer Vision'-gemeenschap gedurende verschillende decennia op brede schaal onderzocht. Om deze taak uit te voeren hebben we verschillende nieuwe beeld-

kenmerken voorgesteld: PPC en BoSP. PPC is een eenvoudig schema dat CNN-kenmerken extraheert en samenneemt vanuit verschillende regio's van het beeld en PCA toepast om de dimensie van de kenmerken te reduceren. BoSP beschouwt de afbeeldingen van de beeld-data op de kenmerken als surrogaat-delen en wijst aan de hand van de activatie-waarden de dichtbevolkte regio's van het beeld toe aan deze surrogaat-delen. Zowel PPC- als BoSP-kenmerken kunnen bepaald worden zonder een significante toename in de berekeningskosten.

Objecten zijn vaak georganiseerd in een hiërarchie. Terwijl de traditionele beeld-classificatie-taak zich enkel richt op de categorieën op het laagste niveau in de uiteinden van de hiërarchie, stellen wij dat de categorieën beter kunnen worden beschreven door de evolutie van de beeld-categorieën. Daarom introduceren we de hiërarchische beeld-classificatie-taak, welke tracht hiërarchische labels van grof naar fijn te genereren in plaats van een enkel label op het laagste niveau. Hiertoe ontwikkelen we een CNN-RNN raamwerk, waarbij de CNN wordt gebruikt om discriminatieve beeldkenmerken te extraheren en de RNN de relatie tussen de hiërarchische categorieën exploiteert om sequentiële labels te genereren. Bovendien onderzoeken we de effectiviteit van het gebruik van dit raamwerk voor de traditionele beeld-classificatie-taak en 'Weakly Supervised Learning'.

Een ander gebruik van het CNN-RNN raamwerk ligt bij de beschrijving van het beeld. Dit is een belangrijke en uitdagende taak in 'Vision-to-Language'-onderzoek. Hierbij is het doel om een beeld te beschrijven met betekenisvolle en zinnige beschrijvingen op het niveau van volledige zinnen. We onderzoeken de effecten van verschillende Convnets op de beeldbeschrijvingen, d.w.z. Single Label Convnet, Multi-Label Convnet en Multi-Attribute Convnet. Alle drie de Convnets richten zich op verschillende delen van de visuele inhoud van het beeld. We stellen voor om ze samen te nemen om zo een rijkere visuele representatie te verkrijgen. Over het algemeen bereiken we competitieve resultaten in vergelijking met de state-of-the-art.