Exploring deep learning for multimodal understanding

Lao, M.

Nederlandse Samenvatting

We leven in een nieuw medialandschap waar de hoeveelheid visuele en taalgegevens exponentieel toeneemt. Daarom is het belangrijk om intelligente multimodale redeneersystemen te ontwerpen om te voldoen aan onze behoefte aan het begrijpen van multimedia gegevens. Een uitstekend voorbeeld van zo'n systeem, dat aanzienlijke aandacht heeft gekregen in de context van vraag-antwoord (QA) scenario's, is een geautomatiseerd Visueel Vraag Antwoord (VQA) systeem. Een VQA-systeem richt zich op het beantwoorden van tekstuele vragen, op basis van afbeeldingen van gebruikers. Dit kan worden toegepast in verschillende praktische scenario's, zoals robotdocenten en slimme huishoudmanagementsystemen. Om de prestaties en robuustheid van multimodale QA-systemen te verbeteren en ze verder in staat te stellen de uitdagingen van de open wereld aan te gaan, onderzoeken we vijf thema's voor efficiënte en praktische VQA-modellen.

Het eerste thema richt zich op efficiënte functiefusie voor visie-taal invoer, wat een cruciale rol speelt bij het verbeteren van de voorspellingsnauwkeurigheid van huidige VQA-modellen. Geïnspireerd door het meerfasige redeneergedrag bij mensen, proberen we meerdere en fijnmazige multimodale interacties te bereiken om de fusieprestaties te verbeteren. Om dit te bereiken introduceren we een nieuw Multi-stage Hybrid Embedding Fusion (MHEF) netwerk om multimodale fusie te verbeteren vanuit het oogpunt van dual-space embedding en meerfasig leren. Met meerfasige MHEF kunnen we niet alleen de prestaties in elke afzonderlijke fase verbeteren, maar ook aanvullende nauwkeurigheidsverbeteringen behalen door alle voorspellingsresultaten van elke fase te integreren.

Het tweede thema van deze thesis is het bestuderen van de taalbias in huidige multimodale redeneeralgoritmen, wat een voornaam obstakel vormt voor modelrobustheid en betrouwbaarheid in de VQA-taak. Ten eerste, om de overmatige afhankelijkheid van dominante antwoorden voor verschillende vraagtypes te verminderen, introduceren we een Taal Prioriteit gebaseerde Focal Loss (LP-Focal Loss) door de standaard kruisentropieverlies in VQA-training te herschalen. Deze geeft dynamisch lagere gewichten aan bevooroordeelde antwoorden bij het berekenen van het trainingsverlies, waardoor de bijdrage van meer bevooroordeelde gevallen in oneevenwichtige trainingsgegevens wordt verminderd. Ten tweede, gebaseerd op het feit dat VQA-voorbeelden met verschillende niveaus van taalbias verschillend bijdragen aan

antwoordvoorspelling, stellen we een nieuw door Taal Bias aangedreven Curriculum Learning (LBCL) framework voor dat de taalbias overwint via een gemakkelijk-naar-moeilijk leeraanpak. Specifiek, in de initiële trainingsfase, leert het VQA-model voornamelijk de oppervlakkige tekstuele correlaties tussen vragen en antwoorden (eenvoudig concept) uit meer bevooroordeelde voorbeelden, en richt zich vervolgens geleidelijk op het leren van het multimodale redeneren (moeilijk concept) uit minder bevooroordeelde voorbeelden in de volgende fasen

In het derde thema richten we ons op het overwegen en analyseren van het biasprobleem in Audio-Visuele Vraag Antwoording (AVQA), wat een complexere vraag-antwoordtaak is die tekstuele-visuele-auditieve informatie betreft voor multimodaal begrip. Door gedetailleerde causale-grafiekanalyses en zorgvuldige inspecties van hun leerprocessen onthullen we dat AVQA-modellen niet alleen geneigd zijn om overheersende taalbias overmatig te benutten, maar ook lijden onder extra gezamenlijke modale biases veroorzaakt door de kortsluitingsrelaties tussen tekstuele-auditieve/visuele samenkomsten en gedomineerde antwoorden. Om dit probleem te verlichten, stellen we een COllabrative CAusal (COCA) Regularisatie voor om dit vanuit twee aspecten te verhelpen. Ten eerste wordt een nieuwe Bias-gecentreerde Causale Regularisatie (BCR) voorgesteld om specifieke kortsluitingsbiases te verminderen door bias-irrelevante causale effecten te interveniëren, en verder de voorspellingen van AVQA-modellen te introspecteren in contrafactuele en feitelijke scenario's. Ten tweede, gebaseerd op het feit dat de overheersende bias die modelrobustheid schaadt voor verschillende voorbeelden de neiging heeft anders te zijn, introduceren we een Multi-shortcut Collaborative Debiasing (MCD) om te meten hoe elk voorbeeld lijdt onder verschillende biases, en dynamisch hun debiasing-concentratie aan te passen aan verschillende kortsluitingscorrelaties.

Het vierde thema heeft betrekking op de uitdagingen van levenslang leren waarmee VQA te maken krijgt in een open wereld, omdat van VQA-systemen altijd wordt verwacht dat ze hun kennis uitbreiden en inspelen op de steeds veranderende behoeften van gebruikers. Daarom introduceren we een nieuwe VQA-instelling genaamd Multi-Domein Levenslang VQA (MDL-VQA). Dit heeft als doel een VQA-systeem te ontwikkelen dat continu kan leren en zich kan aanpassen aan nieuwe domeinen op basis van nieuwe gegevens, terwijl kennis van eerdere domeinen behouden blijft. Om deze uitdagingen het hoofd te bieden, stellen we een nieuw Self-Critical Distillation (SCD) raamwerk voor MDL-VQA voor. Dit verlicht het probleem van vergeten door kennis van het vorige domein over te dragen van leraarmodellen naar leerlingmodellen. Concreet bevat het logits- en kenmerkniveau destillaties om de leraar te stimuleren waardevolle oude kennis over te dragen en tegelijkertijd de leerling te helpen nuttige kennis uit het huidige domein te vergaren. Door deze tweeledige destillaties te combineren, verbetert SCD de stabiliteit van het VQA-model tegen het vergeten, terwijl het zijn vermogen behoudt om nieuwe kennis te leren.

Voor het laatste thema richten we ons op een praktisch gepersonaliseerd federaal leerscenario, waarbij de training van VQA-modellen plaatsvindt over meerdere klanten of servers, terwijl data lokaal blijft, zonder deze te centraliseren. Om het onderzoek te faciliteren, stellen we een FedVQA-instelling voor om gepersonaliseerde VQA-klantmodellen te trainen voor verschillende visuele scènes, terwijl een algemeen model geoptimaliseerd wordt om goed te generaliseren op ongeziene scènes, via klantensamenwerking met inachtneming van de privacybeperking. Om het probleem van het vergeten van algemene kennis in FedVQA aan te pakken, stellen we een nieuw federaal raamwerk voor het behoud van paarsgewijze voorkeuren (FedP^3) voor om gepersonaliseerd leren te verbeteren. Specifiek ontwerpen we eerst een differentieerbare paarsgewijze voorkeur (DPP) om kennisbehoud te verbeteren door een flexibele maar effectieve algemene kennis te formuleren. Vervolgens introduceren we een filter voor vergeten kennis (FKF) om de klantmodellen aan te moedigen selectief makkelijk vergeten kennis te consolideren. Door de DPP en de FKF te combineren, coördineert FedP^3 de algemene en de gepersonaliseerde kennis om het gepersonaliseerde vermogen van klanten en de generaliseerbaarheid van de server te verbeteren.

We hebben uitgebreide experimenten uitgevoerd om de effectiviteit van de voorgestelde benaderingen voor de vijf thema's te verifiëren. De resultaten tonen opmerkelijke verbeteringen ten opzichte van verschillende baselines en de meest geavanceerde methoden. Daarom biedt dit proefschrift belangrijke nieuwe bijdragen, inzichten en bevindingen voor de onderzoeksgemeenschap en toekomstige toepassingen op het gebied van multimodaal begrip en VQA.