Exploring deep learning for intelligent image retrieval

Chen, W.

Nederlandse Samenvatting

We leven in een informatietijdperk. De hoeveelheid beeld- en video-gegevens neemt exponentieel toe. Het is belangrijk om informatiesystemen te ontwikkelen die in staat zijn om zulke grote gegevenscollecties op te slaan, te beheren en te verspreiden. Een van de belangrijkste technieken hierbij is het intelligent ophalen van afbeeldingen; dit is een van de meest onmisbare technieken om te voldoen aan onze behoefte om visuele informatie die van belang is te vinden. Om het mogelijk te maken afbeeldingen intelligent op te halen zijn verschillende algoritmen noodzakelijk, hierbij moet gelet worden op hoge nauwkeurigheid en hoge efficiëntie. Representatiefuncties vormen hierbij de kern van verschillende ophaalalgoritmen die we in dit proefschrift zullen bespreken.

Voor mensen is het eenvoudig om vergelijkbare afbeeldingen uit een verzameling beelden te vinden op basis van een gegeven voorbeeld. Het is echter moeilijk voor een computer om even nauwkeurig te zoeken als mensen vanwege de bestaande semantische kloof tussen de concepten gebruikt door mensen en de beeldkenmerken die gewoonlijk worden afgeleid van afbeeldingen (dwz pixels of symbolen). Bovendien zal het voor de computer moeilijker zijn om nauwkeurig te zoeken als het zoekitem uit verschillende modaliteiten bestaat (*e.g.* tekst, audio *etc*). Dit wordt veroorzaakt door de tweede uitdaging: de heterogeniteitskloof. Deep learning heeft, vooral voor convolutionele neurale netwerken, vooruitgang geboekt bij het aanpakken van deze uitdagingen en het proces van het intelligent ophalen van afbeeldingen aanzienlijk vergemakkelijkt.

Het eerste onderwerp in dit proefschrift is het verkennen van multi-modale retrieval door zowel de visuele als de tekstuele modaliteiten te gebruiken. De moeilijkheid bij dit onderwerp zit hem in de semantische kloof en de heterogeniteitskloof. We ontwerpen een informatie-entropieverliesfunctie op basis van Shannon's informatietheorie om het leren van een gedeelde latente ruimte voor gepaarde beeld- en tekstinvoer te regulariseren. De gebruikelijke praktijk van cross-modale opvraging is om een gedeelde ruimte te construeren waar afbeeldingskenmerken en tekstkenmerken sterk door elkaar worden gehaald, waardoor de overeenkomst tussen afbeelding en tekst verder kan worden geassocieerd. Deze eigenschap van de gedeelde ruimte is consistent met de meting van informatie-entropie volgens Shannon's informatietheorie. Dit idee wordt gedemonstreerd voor cross-modale hashing retrieval waarbij de reële

featurewaarden en de binaire hashcodes worden beperkt/gestuurd door het verlies van informatie-entropie.

Vervolgens integreren we Shannon's informatietheorie en adversarial learning voor cross-modale retrieval. Adversarial learning zorgt voor een betere verdeling van bimodale-kernmerken opdat we de heterogeniteitskloof kunnen overbruggen en zo betere prestaties mogelijk maken. Om de semantische kloof te verkleinen, worden Kullback-Leibler (KL) divergentie en bidirectioneel tripletverlies gebruikt om de intra- en inter-modaliteitsgelijkvormigheid tussen kenmerken in de gemeenschappelijke ruimte te vinden. We ontwerpen ook een regularisatieterm op basis van KL-divergentie met temperatuurschaling om de bevooroordeelde labelclassificatie te kalibreren en zo de onbalans in de basisdata te verminderen.

Het tweede thema van dit proefschrift betreft de vraag hoe we leer-taken van een voorgaande taak kunnen leren, zonder telkens overnieuw te moeten beginnen, of het geleerde te vergeten als nieuwe zaken geleerd worden. We onderscheiden drie belangrijke deelvragen: incrementeel leren voor retrieval in dezelfde fijnmazige dataset, feature-schattingen voor opeenvolgende diepe modellen in incrementeel leren en levenslang leren voor rerieval in verschillende datasets. Voor de eerste deelvraag kijken we naar incrementeel leren voor het vinden van fine-grained afbeeldingen. Dit wordt bereikt door de representatie- en classificatie-distributies te regulariseren. Dit doen we door gebruik te maken van het maximale gemiddelde discrepantieverlies en kennisdestillatieverlies. Om de voorgestelde methode te evalueren, splitsen we een dataset in twee delen, de ene wordt gebruikt als oude data (of oude taken) en de andere wordt gebruikt als de nieuwe data voor incrementele training (of nieuwe taken).

Voor de tweede deelvraag richten we ons op de sequentie van diepe modellen die worden getraind wanneer nieuwe taken opeenvolgend worden toegevoegd. Dit scenario met meerdere taken zal lijden aan catastrofaal vergeten. Het opslaan van de sequentie van modellen voor het overdragen van eerder geleerde kennis is geheugenverslindend. In plaats daarvan stellen we een eenvoudige maar effectieve methode voor het schatten van features voor om deze beperking te verminderen.

Voor de derde deelvraag kijken we naar de praktische kant van levenslang leren voor het zoeken naar beelden, waarbij het neurale network achtereenvolgens wordt getraind op verschillende datasets. De semantische verschuivingen tussen verschillende datasets maken het moeilijk om iets te doen aan het vergeten van geleerde features. We pakken deze beperking aan door een duaal kennisdestillatiekader te gebruiken dat twee professionele supervisors en een instrinsiek-gemotiveerde student omvat. Het ene supervisormodel heeft vaste parameters en wordt gebruikt voor het overdragen van eerder geleerde kennis aan de volgende taken, terwijl een andere on-the-fly supervisor samen met de student wordt opgeleid en verantwoordelijk is voor het overdragen van kennis die is geleerd over de nieuw toegevoegde taken. Verder

gebruiken we ook de statistieken over de BatchNorm-lagen van het bevroren supervisormodel om enkele representatieve afbeeldingen te genereren voor de volgende taken.

Tenslotte voeren we diepgaande experimenten uit om de effectiviteit van de voorgestelde methoden voor de twee onderwerpen vast te stellen. De resultaten laten significante verbeteringen zien ten opzichte van verschillende baselines en state-of-the-art methoden. Dit proefschrift levert nieuwe bijdragen, inzichten en vondsten voor de onderzoeksgemeenschap en toekomstige toepassingen op het gebied van intelligente image-retrieval.