Multi modal representation learning and cross-modal semantic matching

Wang, X.

SAMENVATTING

De mens neemt de wereld waar met zijn zintuigen, te weten: zicht, smaak, gehoor, geur en aanraking. In termen van informatie overdracht beschouwen we deze zintuiglijke waarnemingen als verschillende informatiekanalen of *modalities*. Wanneer we meerdere informatiekanalen tegelijk beschouwen dan spreken we van multi-modale overdracht en de invoer is bekend als multimedia. Multimedia-data zijn van nature complex en bevatten verschillende soorten informatie die met elkaar verweven zijn. We kunnen dit multimodale perspectief gebruiken om betekenis en begrip toe te kennen aan data. Dit is vergelijkbaar met de verwerking van informatie in het menselijke brein, we leren hoe we waarnemingen kunnen combineren en daar zinnige informatie uit halen. In dit proefschrift wordt het leren gedaan met computers en slimme algoritmen. Dit wordt aangeduid met *kunstmatige intelligentie*. Vanuit dat perspectief hebben we, in dit proefschrift, multimedia informatie bestudeerd, met een focus op beeld en tekst voor semantische *mapping*. De doelstellingen van het leren van deze semantische mappings zijn: (1) het leren van word embeddings via afbeeldingen van objecten en relaties; (2) het fijnmazig labellen van objecten in afbeeldingen; (3) *kernel-based data transformation* voor beeld- en tekstassociatie; (4) het leren van beelrepresentaties via een cross-modaal *contrastive learning* framework.

Het eerste doel was het verbeteren van het leren voor de tekstuele representaties van relatiewoorden op basis van visuele supervisie van afbeeldingen. In ons werk stellen we het VS-Word2Vec-model voor om de vectorrepresentatie van relatiewoorden te leren door tegelijkertijd de visuele modaliteit en natuurlijke taal te analysieren. Dit kan de semantische kloof tussen beeld en tekst verkleinen in het vinden van multi-modale informatie. Onze methode kan de visuele *features* van objecten berekenen op basis van analyse van de delen van afbeeldingen die relatiewoorden representeren, en vervolgens de visuele overeenkomstmatrix voor alle relatiewoorden berekenen. Op basis van onze *embedding*-methode krijgen de embeddings van relatiewoorden een betere representatie dan met het originele CBOW-model.

In een beeld kunnen objecten worden herkend en gelabeld als een categorie label, of in een subcategorie van dat label. Als we de subcategorie beschouwen dan spreken we van het aanleren van fijnmazige labels. Het tweede doel was om de vraag te beantwoorden hoe fijnmazige objectlabels bij objectdetectie kunnen worden aangeleerd met behulp van aanvullende informatie gekoppeld aan afbeeldingen. In dit proefschrift stellen we een nieuwe methode voor, genaamd *label inference curriculum network* (LICN), voor het probleem van het fijnmazig leren van objectlabels met *weak supervision* van bijschriften bij afbeeldingen. Met deze methode kan een mapping worden gemaakt op basis van de overeenkomst

tussen de grove categorielabels uit openbare data-sets en de fijnmazige categorielabels die worden geëxtraheerd uit bijschriften. We gebruiken hiervoor een combinatie van embeddingstechnieken en databases. Door deze semantische mapping kunnen objecten beter geïdentificeerd worden met fijnmazige labels en via die labels kunnen betere detectors voor die objecten ontwikkeld worden. Experimentele resultaten op openbare datasets en onze geconstrueerde datasets demonstreren de effectiviteit van onze aanpak en laten zien dat het nuttig is om het trainingsproces te structureren in de volgorde van eenvoudige voorbeelden naar moeilijke voorbeelden. Deze aanpak is bekend als tet zogenaamde curriculum learning framework.

Een volgende aanpak die we hebben bestudeerd is de *kernel-based data transformation* voor beeld- en tekstassociatie. Dit heeft tot doel een probabilistisch *mixture model*, het KMM, te tranen voor het modelleren van de semantische correlatie tussen web-afbeeldingen en tekst. Het KMM leert op basis van de veronderstelling dat de relatie tussen verschillende modaliteiten meerdere basistransformaties volgt, die op een bepaald deel van de afbeelding van toepassing zijn, te weten de regio. Die regio wordt gerepresenteerd door een *neighborhood model* in de vector ruimte. Ons model geeft een oplossing voor de nonlineariteit van de datadistributie en cross-modale mapping via een op kernels gebaseerde theorie. Als een oplossing voor de niet-lineariteitstransformatie voor cross-modale semantische mapping richt ons model zich op de complexiteit van de semantische distributie over de input ruimte, en daarbij de continuïteit ervan op lokale schaal.

Tenslotte hebben we in dit proefschrift het leren van visuele representaties via het cross-modale *contrastive learning* framework bestudeerd. Deze aanpak heeft tot doel een methode te vinden voor cross-modal mapping op basis van *weak supervision* van de bijschriften van afbeeldingen. Het doel is om frases uit afbeeldingsbijschriften te koppelen aan de objecten in de afbeelding. In dit proefschrift hebben we een nieuwe benadering van deze *weakly supervised phrase grounding* voorgesteld op basis van de correspondentie tussen afbeeldingen en bijschriften. Onze belangrijkste bijdrage ligt in het systematisch leren van gecontextualiseerde visuele representaties met een *mixed contrastive loss function.* Met ons model zijn we in staat de best mogelijke nauwkeurigheid te realiseren op de MS COCO- en Flickr30K Entities-testsets.