Expertblog

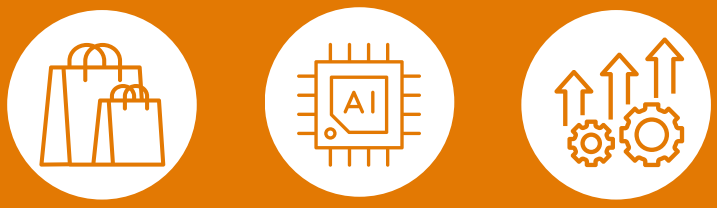
Fine-tuning CLIP voor automatische extractie van productkenmerken uit afbeeldingen

Geschreven door
Britt Deckers

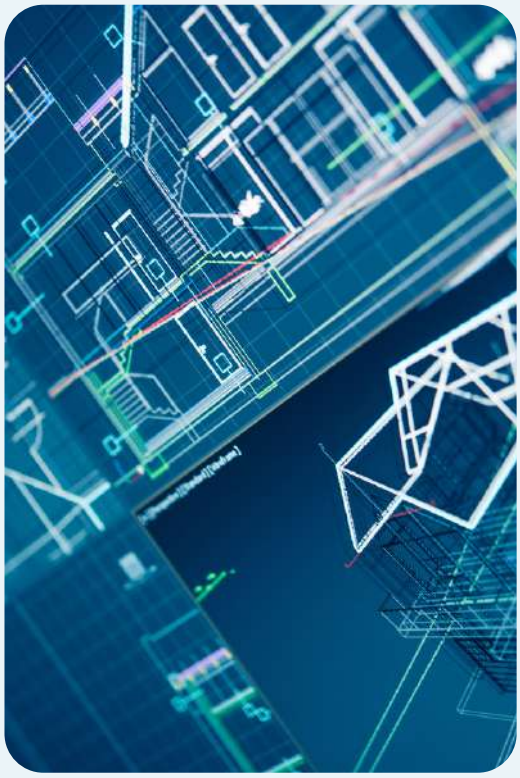
10-03-2023

Introductie

In de afgelopen 10 jaar is de hoeveelheid data in de detailhandel aanzienlijk gegroeid. Winkeliers verzamelen nu een breed scala aan datapunten over de producten die ze verkopen om klantinformatie te verbeteren, de datakwaliteit te verbeteren en producten op kopers af te stemmen. De kwaliteit en kwantiteit van deze productgegevens kan echter sterk variëren tussen bedrijven, waarbij sommige bedrijven over voldoende gegevens van hoge kwaliteit beschikken, terwijl andere aanzienlijke hiaten vertonen. Sommige detailhandelaren hebben bijvoorbeeld alleen productafbeeldingen en geen lijst met productkenmerken (bijv. kleur = rood). Het extraheren van deze functies kan nuttig zijn voor verschillende doeleinden, zoals gegevensverrijking, detectie van problemen met gegevenskwaliteit en verbetering van gegevenskwaliteit. De afgelopen jaren zijn er aanzienlijke vorderingen gemaakt op het gebied van taal- en visiemodellen. Deze multimodale modellen, zoals CLIP van OpenAI (Radford et al., 2021), kunnen de kloof tussen tekst en afbeeldingen overbruggen door overeenkomsten tussen de tekstuele en visuele ruimten te identificeren. In dit artikel wordt een methode beschreven voor het afstemmen van CLIP voor het automatisch extraheren van productkenmerken uit afbeeldingen.



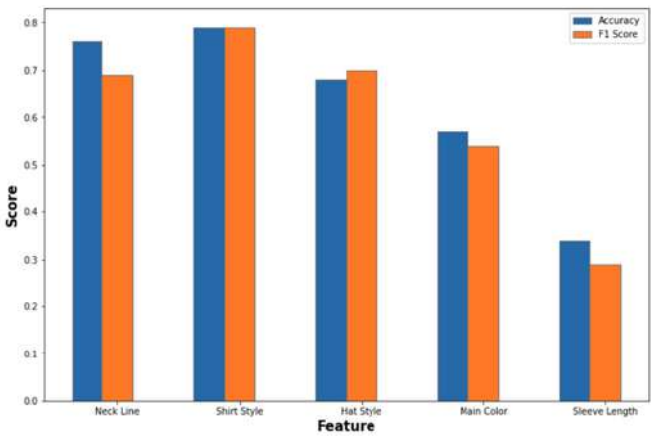
Inleiding tot CLIP



CLIP (Contrastive Language-Image Pre-training) is een model dat is ontwikkeld om taal te verbinden met visie. Het is een groot multimodaal neuraal netwerk dat is getraind op een grote verscheidenheid aan afbeeldingen die zijn gelabeld met natuurlijke taal om de relatie tussen natuurlijke taal en visuele objecten te begrijpen. CLIP wordt getraind met behulp van een proces dat contrastief leren wordt genoemd. In dit proces krijgt het model een reeks positieve en negatieve voorbeelden voorgeschoteld en leert het onderscheid te maken tussen de twee. In het geval van CLIP bestaan de positieve voorbeelden uit paren afbeeldingen en tekstbijschriften die de afbeeldingen beschrijven, terwijl de negatieve voorbeelden bestaan uit paren afbeeldingen en tekstbijschriften die niet overeenkomen. Tijdens de training krijgt het model een afbeelding en een tekstbijschrift te zien, en het moet bepalen of het tekstbijschrift de afbeelding al dan niet nauwkeurig beschrijft. Het model is getraind om de overeenkomst tussen de afbeelding en het tekstbijschrift te maximaliseren wanneer ze overeenkomen en de overeenkomst te minimaliseren wanneer ze niet overeenkomen. CLIP combineert de sterke punten van zowel beeldherkenningsmodellen als taalmodellen en is in staat om te generaliseren naar nieuwe objecten en scènes. CLIP kan worden gebruikt door natuurlijke taalprompts te formuleren om een breed scala aan taken uit te voeren, zoals beeldonderschriften, objectdetectie en visuele vraagbeantwoording.

CLIP voor het extraheren van productkenmerken uit afbeeldingen

Voor de taak die voorhanden is, het extraheren van productkenmerken uit afbeeldingen, wordt een afbeelding aan het model gepresenteerd met een lijst met mogelijke bijschriften, bijvoorbeeld ["Een foto van een top met lange mouwen", "Een foto van een top met korte mouwen" , "Een foto van een mouwloos topje"], en het model voert een lijst met waarschijnlijkheden uit voor elk gegeven bijschrift. Het is belangrijk dat de prompts op dezelfde manier worden geformuleerd als de bijschriften in de trainingsdataset. Zhou et al. (2022) laten zien dat het belangrijk is om context aan de prompt toe te voegen, daarom wordt er extra informatie aan de prompt toegevoegd, en niet alleen het kenmerk. Ze laten zien dat het afstemmen van woorden op een specifiek domein een enorme impact kan hebben op de prestaties van deze modellen. Om de prestaties van standaard CLIP voor het extraheren van beeldkenmerken te testen, is een dataset werd gebruikt dat afbeeldingen bevat van kleding met hun kenmerken, zoals kleur, mouwlengte, halslijn, enz. CLIP is geïmplementeerd en getest op meerdere kenmerken van de gegevens.



CLIP afstellen

TOm de prestaties van CLIP bij het extraheren van productkenmerken te verbeteren, hebben we CLIP verfijnd voor het domein van productafbeeldingen. Om CLIP te verfijnen, werden er meerdere tests gedaan op de bovengenoemde kledingdataset om de trainingslus te perfectioneren. Eerst werden de gegevens opgesplitst in trein-, validatie- en testsets en werd een gebalanceerde batch-sampler geïmplementeerd om ervoor te zorgen dat de verschillende kenmerkwaarden voor elke batch in evenwicht zijn om vertekening na training te voorkomen. Er werd een cross-entropieverlies gebruikt en het model met het kleinste verlies, berekend op de validatieset tijdens de training, van alle tijdperken wordt opgeslagen als het beste model. Het getrainde model is vervolgens getest op de testset om te zien of het finetunen van CLIP de scores van het model daadwerkelijk verbetert. De cijfers tonen de hoeveelheid trainingsgegevens die nodig zijn om CLIP te verfijnen om behoorlijke nauwkeurigheid en F1-scores te bereiken. Voor de kledinggegevensset laten de cijfers zien dat er veel trainingsgegevens nodig zijn om een nauwkeurigheidsscore van ~0,7 te bereiken. Dit lijkt enigszins inefficiënt, aangezien u in het begin dus veel gelabelde gegevens nodig zou hebben. Aangezien het zien van bepaalde kenmerken op kledingstukken zelfs dubbelzinnig kan zijn voor het menselijk oog, er zijn bijvoorbeeld vaak meerdere kleuren, wat het moeilijker maakt om een hoofdkleur te kiezen, wordt verondersteld dat CLIP minder trainingsgegevens nodig heeft wanneer de kenmerken duidelijker zijn zichtbaar. Deze hypothese is getest op een dataset met potten en pannen, waarbij de gewenste kenmerken waren of de pannen een deksel hebben of niet en of ze een handvat hebben of niet. De resultaten worden weergegeven in figuur 1. Voor het nauwkeurig afstemmen van CLIP op deze functies waren aanzienlijk minder trainingsgegevens nodig, met ~60 trainingsgegevenspunten, kon een nauwkeurigheid en F1-score van ~1 worden bereikt.

Zoals eerder vermeld, is de formulering van de prompts belangrijk voor het verkrijgen van een nauwkeurige uitkomst van modellen zoals CLIP. Daarom wordt verondersteld dat dit ook betere resultaten zou opleveren voor het trainen van CLIP op onderschriften in natuurlijke taal die rond de functielabels worden gegenereerd. Na het uitvoeren van tests op het trainen van CLIP met meerdere vormen van ondertiteling die de functie bevatten, werd echter geconcludeerd dat de resultaten voor fijnafstemming beter zijn wanneer de ondertiteling alleen uit de gewenste functie bestaat. Dus in plaats van een afbeelding te trainen met de tekst "Een foto van een top met lange mouwen", werd de afbeelding getraind op "Longsleeved".

Figure 1

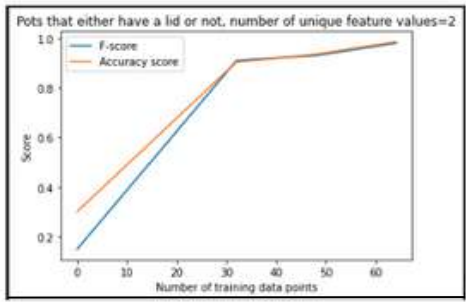
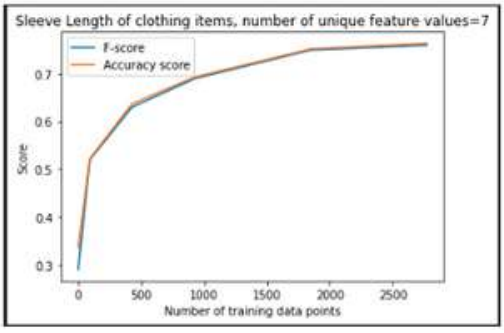


Figure 2



Conclusie

Concluderend kan CLIP van OpenAI worden gebruikt voor het automatisch extraheren van productkenmerken uit afbeeldingen. Voor veel productcategorieën en functies is deze taak echter te domeinspecifiek en moet CLIP nauwkeurig worden afgesteld om redelijke resultaten te krijgen. Daarom biedt dit artikel een leidraad voor hoe CLIP kan worden verfijnd voor het extraheren van productkenmerken uit afbeeldingen. Hiervoor heb je een set gelabelde afbeeldingen nodig, de grootte van deze set hangt af van hoe zichtbaar de kenmerken zijn op je afbeeldingen. Hoe duidelijker zichtbaar, hoe minder trainingsgegevens er nodig zijn.

Contact Squadra MLC