Generative Adversarial Networks for Diverse and Explainable Text-to-Image Generation

Zhang, Zhenxing

DOI:
10.33612/diss.507581028

NEDERLANDSE SAMENVATTING

Dit proefschrift richt zich op het bestuderen van de AI-taak 'tekst-naar-beeld synthese', die gericht is op het opleveren van fotorealistische en semantisch consistente afbeeldingen waneer er een tekstuele beschrijving als invoer is gegeven. Tekst-naar-beeldsynthese heeft steeds meer aandacht gekregen op het gebied van 'computer vision' en natuurlijke taalverwerking binnen de AI. Immers, mensen zijn in staat om zich een visuele voorstelling te maken op basis van een korte talige beschrijving van een voorwerp of een scene. Als we echte kunstmatige intelligentie willen ontwikkelen die ons begrijpt en die ons kan ondersteunen in onze behoefte aan visuele informatie en afbeeldingen dan zullen we ook algoritmes voor visueel voorstellingsvermogen moeten ontwikkelen. "Wat ik niet kan creëren, begrijp ik niet", zei de beroemde natuurkundige Richard Feynman. In de geest van dit citaat gaan we ervan uit dat een machine uiteindelijk (belangrijke aspecten van) visuele inhoud zal begrijpen als deze in staat is om getrouwe beelden te produceren op basis van een tekstuele beschrijving. Het grote voordeel is dat gegenereerde afbeeldingen door mensen kunnen worden getoetst aan de corresponderende tekst om te beoordelen of het klopt. Behalve een theoretische uitdaging binnen het AI onderzoek zijn er ook nuttige praktische toepassingen. Voorbeelden zijn: het snel en gemakkelijk produceren van afbeeldingen van hoge kwaliteit op basis van ingetikte tekst; verfijning van tekstuele zoekopdrachten naar producten op internet door te testen of de genoemde tekstelementen in de gegenereerde afbeelding voorkomen; meubelontwerp, het produceren van digitale kunst, het uitbreiden van visuele trainingsdata voor beeldherkenningsalgoritmes, kunstcreatie, gegevensuitbreiding voor het trainen van beeldclassificatie-algoritmes, visuele controle van de uitvoer van ondertitelings- of vertalingsalgoritmen; fotobewerking volgens tekstuele beschrijvingen, etc.

Er is opmerkelijke vooruitgang geboekt in deze onderzoeksrichting, gebruikmakend van met name het conditioneel-generatieve 'adversarial network' (cGAN). Een op cGAN gebaseerd tekst-naar-beeld-synthesemodel staat echter nog steeds voor meerdere uitdagingen bij het genereren van visueel aantrekkelijke en semantisch consistente voorbeelden. De toepasbaarheid in de praktijk wordt belemmerd door: (1) de moeilijkheid om de tekstgeconditioneerde tekst met hoge resolutie direct te realiseren met behulp van een enkel netwerk met een generator/discriminatorpaar; (2) het frequente optreden van een gebrek aan diversiteit als er meerdere willekeurige afbeeldingen worden geproduceerd op basis van een bepaald stuk tekst; (3) de hoge kans dat veel onsuccesvolle patronen worden gesynthetiseerd bij de willekeurige trekking van een patroon; en (4), de moeilijke verklaarbaarheid van het effect van de conditionele tekstelementen in het cGAN-framework. Om

deze uitdagingen aan te pakken, worden in dit proefschrift uitgebreid studie gedaan naar ontwerpvarianten van nieuwe neurale-netwerkarchitecturen, met als doel het vergroten van de kans op een succescvolle trekking van aanvaardbare, geloofwaardige steekproeven. Het verbeteren van de verklaarbaarheid van conditionele tekst-naar-beeld GAN-modellen wordt aangepakt zowel vanuit het perspectief van de latente ruimte voor de beeldgeneratie (de visuele 'embedding') alsook de linguïstische ruimte (de tekstuele 'embedding'). Dit proefschrift omvat zeven hoofdstukken: Hoofdstuk 1 geeft een korte algemene inleiding tot het onderzoek naar beeldsynthese op basis van taalkundige (tekstuele) beschrijvingen. Verder worden de onderzoeksvragen van deze studie en de belangrijkste nieuwe resultaten en bijdragen toegelicht.

Hoofdstuk 2 is gericht op algoritmes voor het produceren van perceptueel plausibele afbeeldingen die overeenkomen met de gegeven natuurlijke taalbeschrijvingen, waarbij slechts gebruik wordt gemaakt van een enkel generator/discriminator-paar, bestaande uit een enkel netwerk. Hiertoe stellen we het Dual-Attention Generative-Adversarial Network (DTGAN) voor voor tekst-naar-beeldsynthese. In het bijzonder introduceert DTGAN kanaal-(=kleur)bewuste en pixel-(=vorm) bewuste aandachtsmodules die de generator kunnen sturen om meer te focussen op tekstrelevante kenmerkskanalen en pixelwaarden door de aandachtsgewichten te berekenen tussen de globale zinsvector en de twee bovengenoemde factoren kleur en vorm. Belangrijker is dat DTGAN de tekstgebaseerde aandacht verspreidt over vele, zelfs alle lagen van het generatornetwerk. Dit zorgt voor invloed van de tekst op kenmerken op alle verschillende hiërarchische niveaus in de pijplijnarchitectuur, van ruwe, vroege kenmerken tot abstracte late kenmerken. Daarna hebben we de rekenstap Conditional Adaptive Instance-Layer Normalization (CAdaILN) ontwikkeld om de taalkundige aanwijzingen van de zinsinsluiting in staat te stellen flexibel de hoeveelheid verandering in vorm en textuur te manipuleren, de visueel-semantische representatie verder te verbeteren en de training te helpen stabiliseren. Vervolgens wordt een nieuw type visuele verliesfunctie ('loss') gebruikt om de beeldresolutie te verbeteren door te zorgen voor levendige vormen en perceptueel uniforme kleurverdelingen van gegenereerde afbeeldingen. De vier bovengenoemde nieuwe componenten dragen bij aan een veelbelovend *ééntraps* tekst-naar-beeld syntheseraamwerk. De resultaten op standaard benchmarkgegevenssets tonen de superioriteit van onze voorgestelde methode aan in vergelijking met de state-of-the-art modellen met een traditionele meertrapsarchitectuur. Hoofdstuk 3 heeft als doel om het probleem aan te pakken van het gebrek aan diversiteit in gegenereerde afbeeldingen voor een bepaalde tekst, in de bestaande *ééntraps* tekst-naar-beeld generatiemodellen bij het toevoegen van willekeurige ruis op de stuurvector. Het lijkt er hierbij op of bepaalde elementen uit de trainingsdata nog steeds zichtbaar zijn en vaak herhaald worden. Om dit probleem aan te pakken, presenteren we een efficiënt en effectief ééntraps raamwerk (DiverGAN) om diverse, fotorealistische en semantisch gerelateerde afbeeldingen te produceren op basis van een tekstbeschrijving

en verschillende willekeurige, latente stuurvectoren. DiverGAN ontwikkelt twee nieuwe attentiemodules op woordniveau, d.w.z. een kanaal-attentiemodule en een pixel-attentiemodule. De kanalen zullen in het algemeen de kleuren zijn (R,G,B) maar het principe is breder toepasbaar. De attentiemodules kunnen het belang van elk woord in de gegeven tekstbeschrijving modelleren, terwijl het netwerk verschillende gewichtswaarden kan toekennen aan de significante kanalen en pixels die semantisch aansluiten bij de meest opvallende woorden. Met behulp van deze twee woordgerelateerde modulatiemodules is DiverGAN in staat om de attributen van de tekstbeschrijving effectief te ontwarren en de lokale weergave van onderdelen in synthetische afbeeldingen nauwkeuriger te bepalen. Daarna is een 'dual-residual architecture' toegevoegd om visuele kenmerken effectief te propageren en tegelijkertijd diepere netwerken mogelijk te maken. Dit resulteert in een hogere convergentiesnelheid en meer levendige details. De kern van de oplossing van het probleem van gebrek aan diversitie is aangepakt door midden in het netwerk een volledig verbonden laag in de pijplijn aan te brengen, om te voorkomen dat beeldelementen al te gemakkelijk doorsijpelen naar de output. Het netwerk wordt namelijk gedwongen de twee-dimensionale afbeelding door een één-dimensionale vector te persen, waardoor het generatieve vermogen van het netwerk opmerkelijk verbeterde. Er treedt een betere balancering op tussen enerzijds de laagdimensionale, willekeurige latente code die bijdraagt aan variatie en diversiteit en anderzijds de modulatiemodules die gebruik maken van hoogdimensionale en tekstuele contexten om de 'feature maps' te beinvloeden. Door een lineaire laag aan te brengen na het tweede *residual block* wordt de beste variëteit en kwaliteit bereikt. Zowel de kwalitatieve als kwantitatieve resultaten op standaard benchmark-data sets ondersteunen de effectiviteit van DiverGAN voor het realiseren van diversiteit, zonder afbreuk te doen aan de kwaliteit en semantische consistentie. In hoofdstuk 4 proberen we ervoor te zorgen dat de kans vergroot wordt dat gegenereerde afbeeldingen geloofwaardig of natuurlijk zijn. Hiervoor analyseren we de hoogdimensionale ruimte van de latente stuurvectors. We doen dit door punten te kiezen die tekstueel (semantisch) duidelijk zijn, en bekijken dan middels lineaire interpolatie wat er gebeurt met de afbeeldingen als je een traject aflegt van een succesvolle latente startpuntcode en een niet-succesvolle latente eindpuntvector. We ontdekten dat het eerste deel van de interpolatieresultaten meestal realistisch is. Voorbij een omslagpunt worden de afbeeldingen in het laatste deel van het traject echter niet meer geloofwaardig. We gaan er van uit dat dit omslagpunt berekend kan worden. Op basis van deze aanname hebben we twee nieuwe data sets geconstrueerd (d.w.z. de Good & Bad data sets, voor zowel afbeeldingen van vogels als van gezichten). De 'goede' en 'slechte' voorbeelden zijn geselecteerd volgens strikte criteria. Om afbeeldingen van hoge kwaliteit te verkrijgen kunnen we nu de kans op het genereren van goede latente codes vergroten door gebruik van een speciale 'Good/Bad classifier'. De resultaten op de ontworpen DiverGAN-generator geven aan dat onze classifier een nauwkeurigheid van meer dan 98% behaalt bij het detecteren van

goede/slechte gegenereerde afbeeldingen, voordat deze aan gebruikers worden aangeboden. Onze dataset is beschikbaar op Zenodo[1].

In Hoofdstuk 5 wordt de relatie tussen de latente controleruimte en de verkregen beeldvariatie onderzocht. Hiertoe analyseren we het generatiemechanisme van een voorwaardelijk tekst-naar-beeld GAN-model (cGAN). Het is mogelijk om betekenisvolle (semantisch relevante) factoren te ontdekken door middel van onafhankelijke componentenanalyse (ICA) en het gebruik van een extra orthogonaliteitsbeperking. Hierdoor zijn de ontdekte richtingen zijn niet alleen onafhankelijk, maar ook orthogonaal. Verder ontwikkelen we een verliesfunctie, 'background-flattening loss' (BFL), om de achtergrondweergave in de afbeeldingen te verbeteren. We analyseren ook wiskundig de overeenkomsten tussen verschillende methodes, Semantic Factorization (SeFa), GANSpace en reguliere PCA. Onze gepresenteerde methode is niet alleen in staat om verschillende interpreteerbare latente-ruimterichtingen af te leiden, maar ook een nauwkeurigere controle te bieden over de latente ruimte dan de traditionele methode van principale componentenanalyse (PCA).

Hoofdstuk 6 heeft tot doel een diepere blik te werpen in de tekstuele (talige) ruimte van een cGAN model om een betere verklaarbaarheid van het algoritme te verkrijgen, ook voor de gebruikers die afbeeldingen willen genereren. We doen dit door een kwalitative analyse van de invloed van de 'linguïstische' inbeddingen. Door middel van lineaire interpolatie tussen deze tekstgebaseerde vectoren van 'sleutelparen' kunnen we zien hoe een afbeelding veranderd, als je een traject langs de onderliggende talige dimensie doorloopt. Meer specifiek visualiseren we de afbeeldingen die zijn gegenereerd door twee contrastieve trefwoorden op de eindpunten neer te zetten, bijvoorbeeld de kleur, de grootte, de lengte van de snavel op de CUB-vogeldataset en de achtergrond, het object op de voorgrond, de handeling (werkwoord), in de MSCOCO-dataset. Door de interpolatie kunnen we het langzame overgangsproces van het eerste beeld naar het resultaat op het andere uiteinde van de as observeren. Dit principe konden we zelfs uitbreiden naar interpolatie in een driehoekig veld, nl. tussen drie tekstuele ankerpunten. Dit is gerealiseerd met een 2-simplex, d.w.z. een triplet-contrast. Het bleek dat de variatie van de afbeeldingen in zo'n driehoek begrijpelijk en vloeiend was. De resultaten konden worden toegepast op onze eerder geintroduceerde DiverGAN methode en werd getest op twee benchmark data sets. In een toepassing zouden gebruikers met deze methode meer gericht naar goede genereerde afbeeldingen kunnen zoeken, in plaats van op goed geluk telkens nieuwe afbeeldingen te genereren.

Hoofdstuk 7 vat de belangrijkste bijdragen van dit proefschrift samen en geeft antwoorden op de onderzoeksvragen. Daarnaast worden een aantal mogelijke nieuwe onderzoeksrichtingen die samenhangen met dit proefschrift besproken.

1 https://zenodo.org/record/6283798#.YhkN_ujMI2w

LIST OF PUBLICATIONS BY THE AUTHOR

1. Z. Zhang and L. Schomaker. "DTGAN: Dual attention generative adversarial networks for text-to-image generation." 2021 International Joint Conference on Neural Networks (IJCNN), IEEE, 2021, pp. 1–8.

2. Z. Zhang and L. Schomaker. "DiverGAN: An efficient and effective single-stage framework for diverse text-to-image generation." Neurocomputing, vol. 473, pp. 182–198, 2022.

3. Z. Zhang and L. Schomaker. "Optimized latent-code selection for explainable conditional text-to-image GANs." 2022 International Joint Conference on Neural Networks (IJCNN), IEEE, 2022.

4. Z. Zhang and L. Schomaker. "OptGAN: Optimizing and Interpreting the Latent Space of the Conditional Text-to-Image GANs." arXiv preprint arXiv:2202.12929, 2022.

5. Z. Zhang and L. Schomaker. "Fusion-S2iGan: An Efficient and Effective Single-Stage Framework for Speech-to-Image Generation." under review.