Geluid-naar-beeldsynthese in de webbrowser

Mick Broer

Mei, 2023

Inleiding

Met de opkomst van verschillende AI instrumenten, die tekst weten te converteren naar beeld, is beeldsynthese steeds populairder geworden in de wereld van generatieve kunst. Tekst is niet het enige type input die je kan gebruiken om beeld te genereren. Tegenwoordig kunnen we namelijk ook geluid gebruiken om beeld te synthetiseren, geluid-naar-beeldsynthese noemen we dit. We zullen enkele vertaalmethodes en hoe we deze kunnen implementeren in de webbrowser behandelen. Daarnaast besteden we ook aandacht aan een aantal creatieve systemen die gebruik maken van geluid-naar-beeldsynthese. We zullen een introductie en overzicht van het samenvoegen van beeldsynthese en audio behandelen, vanuit een artistiek perspectief. Geluid-naar-beeldsynthese kent ook veel praktische toepassingen, maar we zullen ons beperken tot de creatieve toepassingen van dit proces.

Definitie

Geluid-naar-beeldsynthese is een proces waarbij afbeeldingen worden gegenereerd met behulp van computermodellen en -algoritmes, op basis van de analyse van een geluid. Het visualiseren van geluid wordt hier in de handen van het ontworpen systeem gelegd. De meest geavanceerde en populaire tool voor beeldsynthese is op het moment DALL·E 2, van OpenAI. Andere populaire instrumenten voor beeldsynthese zijn Stable Diffusion en Midjourney.
Geluid-naar-beeldsynthese is een vorm van cross modal image synthesis. Dit is een vorm van beeldsynthese waarbij beeld wordt gegenereerd vanuit een andere modaliteit, zoals tekstuele beschrijvingen of schetsen. In het geval van geluid-naar-beeldsynthese is dat dus geluid. Er is geen directe relatie tussen beeld- en geluidsdata, wat betekent dat je een creatieve vertaalmethode moet gebruiken om de conversie te maken[1]. De vertaalmethode is onderdeel van het creatieve proces en om deze reden niet te scheiden van het creatieve resultaat. De vertaalmethode is dan ook een belangrijke esthetische parameter van het werk.

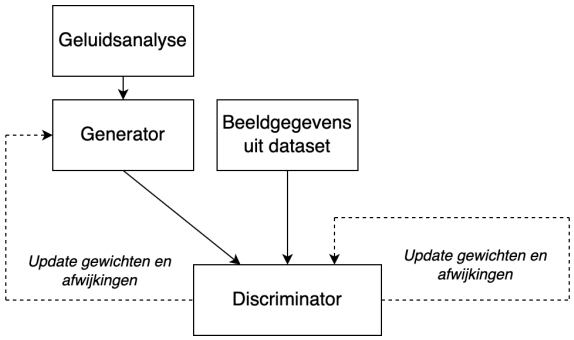
Conversieproces

Het genereren van beeld vanuit geluid kun je verdelen in drie deelprocessen:

- geluidsanalyse
- dataconversie
- beeldsynthese

Het type geluidsanalyse die je maakt is afhankelijk van je doel en voorzieningen. Er moet altijd een afweging gemaakt worden tussen efficiëntie en nauwkeurigheid van de analyse. Aangezien de webbrowser beperkte toegang heeft tot het computergeheugen kan het verstandig zijn om de analyse zo efficiënt mogelijk te maken. Een veelgebruikte methode voor de geluidsanalyse is het gebruik van Mel Frequentie Cepstrale Coëfficiënten (MFCC), omdat deze methode minder waarden oplevert dan bijvoorbeeld een Fourier-transformatie, maar toch erg nauwkeurig is.
Dataconversie is een creatief proces. Het is namelijk

[1] Marije Baalman "Composing Interactions, An Artist's guide to Building Expressive Interactive Systems" chapter 22 (2022)



Figuur 1: *Voorbeeld van een GAN die het mogelijk maakt om geluid te vertalen naar beeld.*

Figuur 2: *Lineair systeem dat geluid vertaalt naar tekst om die tekst vervolgens te converteren naar beeld.*

esthetisch niet interessant om een één op één visuele weergave te hebben van geluid.

We hebben al veel tools die nauwkeurige spectrogrammen en waveform-representaties weten te maken van ons geluid. Er is geen direct verband tussen geluid en beeld dus is het aan de kunstenaar om een creatieve conversie te maken tussen de twee modaliteiten.
Een methode die veel gebruikt wordt om geluid te vertalen naar beeld is het toepassen van een Generative Adversarial Network (zie fig. 1). Dit is een deep learning model dat bestaat uit twee neurale netwerken: een generator en een discriminator [2]. De generator ontvangt in dit geval audiodata, die hij interpreteert om een nieuw beeld te synthetiseren. De discriminator bekijkt het gegenereerde beeld en bepaalt of het gegenereerd is of niet. In het trainingsproces, dat adversarial training heet, probeert de generator een beeld te genereren dat zo goed mogelijk klopt bij de data vergaard uit de audio analyse [3].

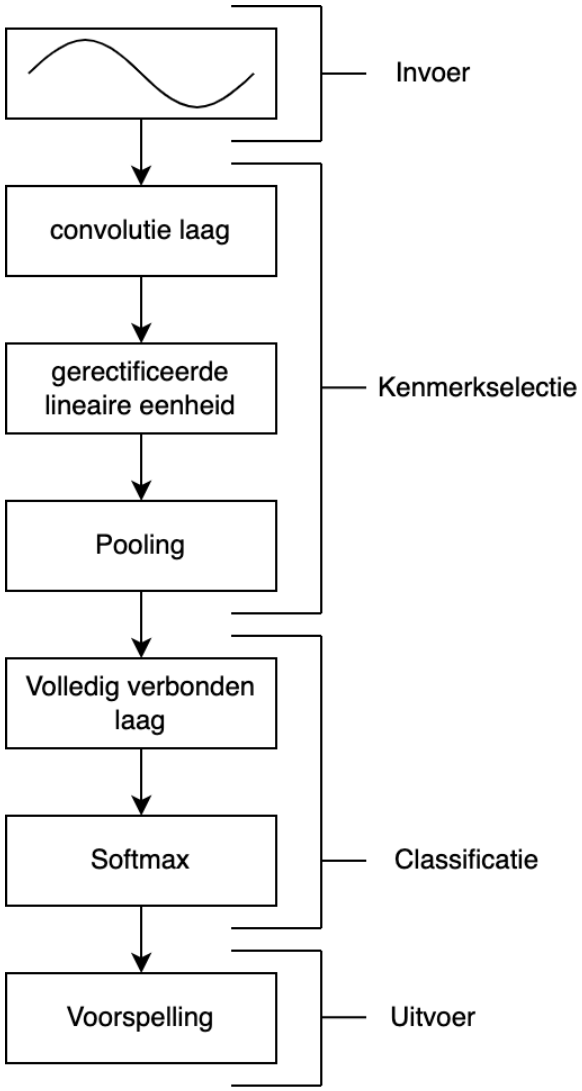
Een andere methode om geluid te vertalen naar beeld is door het geluid eerst te vertalen naar tekst voordat je deze converteert naar beeld (zie fig. 2). De grote creatieve parameter bij deze methode is de woordkeuze. Welke taal gebruik je om het geluid te beschrijven op basis van de geluidsanalyse? De taal die je kiest is namelijk toonaangevend voor het beeld. Om deze methode toe te passen moet je een dataset samenstellen, van paren van geluidsbestanden en hun classificaties. Bij de classificaties is het creatief interessant om technische beschrijvingen van het geluidsbestand te vermijden. Wellicht klinkt een geluidsbestand wel als een boom, of als gisteravond, in plaats van als een sinus van 440 Hertz. Als je voldoende trainingsdata hebt verzameld, kun je een systeem trainen op deze data, zodat het beschrijvingen kan voorspellen op basis van nieuwe geluidsanalyses.

Een veelvoorkomend type netwerk dat wordt gebruikt voor geluidsclassificatie is een convolutioneel neuraal netwerk (zie fig. 3). Dit type netwerk is oorspronkelijk ontworpen voor beeldclassificatie, maar kan ook effectief worden toegepast op geluidsclassificatie. In plaats van 2D-afbeeldingen als input, kan het convolutioneel neuraal netwerk worden getraind met behulp van spectrogrammen of andere tijd-frequentie weergaven van de audio-input, die het netwerk kan gebruiken om verschillende kenmerken van het geluid te leren en te classificeren.

Wanneer het neuraal netwerk beschrijvingen van het geluid voorspelt, is het mogelijk deze beschrijvingen te voeden aan één van de vele tools die tekst-naar-beeld conversie mogelijk maken. Er zijn diverse API's die je kunt gebruiken voor deze taak, zoals de OpenAI API, DeepAI API en ClipDrop API.

[2] Chia-Hung Wan, Shun-Po Chuang, Hung-Yi Lee "Towards Audio to Scene Image Synthesis using Generative Adversarial Network."(2018)

[3] He Huang, Philip S. Yu and Changhu Wang "An Introduction to Image Synthesis with Generative Adversarial Nets"(2018)



Figuur 3: *Voorbeeld van de architectuur van een convolutioneel neuraal netwerk.*

Implementatie in de browser

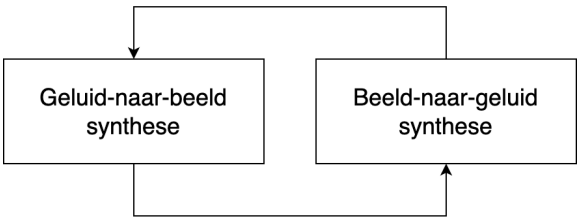
In 2023 is geluid-naar-beeldsynthese nog een relatief traag proces om in de webbrowser toe te passen. Browsers hebben namelijk beperkte toegang tot het besturingssysteem en de hardwarebronnen van de computer, terwijl dit proces veel rekenkracht en geheugen vereist.
Een belangrijk voordeel van geluid-naar-beeldsynthese in de webbrowser toepassen, is dat het toegankelijker wordt voor anderen om je werk te bekijken en er mee te interacteren. Mensen hoeven slechts naar je website te navigeren om er gebruik van te maken. Het maken van aanpassingen aan de site zal ook relatief soepel verlopen, de gebruiker hoeft namelijk geen updates te installeren om een up-to-date versie van het werk te zien.

Voor het analyseren van geluid kun je de Web Audio API gebruiken. Dit is een API die standaard in de JavaScript library zit en door vrijwel alle browsers wordt ondersteund. Deze API is in staat een Fast Fourier-Transform uit te voeren, met haar AnalyserNode. De Web Audio API heeft alleen geen standaard analyser die MFCC waarden kan uitlezen, maar je zou deze waarden zelf kunnen berekenen op basis van de Fast Fourier-Transform. Externe libraries als Meyda.js hebben wel ingebouwde functies die deze berekeningen voor je kunnen doen[4].

Voor het uitvoeren van de dataconversie zijn TensorFlow.js, ML5.js en Brain.js goede opties, ze kunnen namelijk (in real-time) machine learning algoritmes draaien. Deze frameworks kunnen gebruikt worden voor het trainen van een neuraal netwerk vanaf het absolute begin, maar ze kunnen ook voorgetrainde modellen gebruiken. Dit betekent dat het mogelijk is om een model in de browser te gebruiken, dat je hebt getraind op een efficiëntere locatie dan de browser, zoals een externe server.

Het is mogelijk om gebruik te maken van een statische website om je modellen te draaien, maar wanneer je gebruik maakt van zwaardere modellen

[4]Hugh Rawlinson, Nevo Segal, Jakub Fiala "Meyda: an audio feature extraction library for the Web Audio API"(2014)



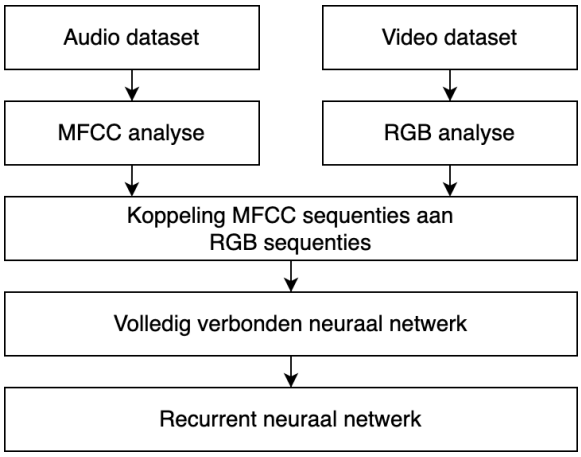
Figuur 4: *Generatief feedback systeem dat gebruik maakt van geluid-naar-beeldsynthese.*

voor je website, is het slim om deze modellen te laten draaien op een server, waar je met je eigen website, als client, mee interacteert. Deze aanpak kan de efficiëntie van je website vergroten en tevens de creatieve mogelijkheden vergroten. De interactie van iedere gebruiker kan opgeslagen worden in het geheugen van de server, hierdoor is het bijvoorbeeld mogelijk om je systeem een extra interactielaag te geven door deze te trainen op de interactie van iedere gebruiker.

Creatieve systemen

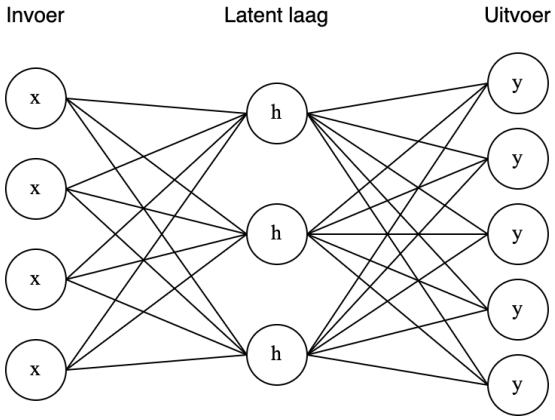
Geluid-naar-beeldsynthese kan een bouwsteen in een groter, complexer systeem zijn. Een systeem dat geluid naar beeld kan vertalen kan erg goed werken in combinatie met een systeem dat beeld naar geluid kan converteren (zie fig. 4). Hier wordt het beeld dat het resultaat is van de geluid-naar-beeldsynthese gevoed in een systeem dat beeld naar geluid kan vertalen. Deze uitvoer, het audiosignaal, laat je daarna weer naar beeld vertalen. Vervolgens begin het proces weer van voor af aan. Hierdoor maak je een feedback lus, waardoor je onvoorspelbaar nieuw materiaal kunt creëren, afhankelijk van de trainingsdata.

Een andere interessante toepassing is geluid-naar-beeldsynthese inzetten als reactieve visualisatie van je muziek. Het is namelijk ook mogelijk om deze techniek te gebruiken om een opeenvolging aan beelden te genereren op basis van geluid, in plaats van slechts één beeld. Het trainen van een systeem dat in staat is deze taak uit te voeren zal echter complexer zijn dan een systeem dat gemaakt is om slechts een enkel beeld te genereren vanuit de audioanalyse (zie fig. 5).

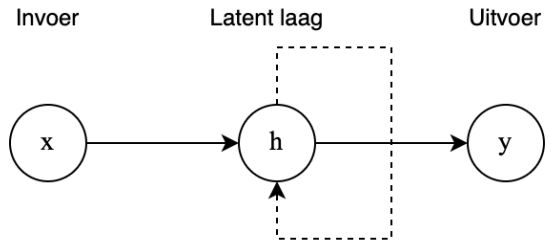


Figuur 5: *Verloop van het trainen van een systeem voor audio-naar-video conversie.*

Het systeem moet leren omgaan met zowel audio materiaal, als sequenties van afbeeldingen, het gegenereerde materiaal bevindt zich namelijk in het tijdsdomein. Bij het maken van dit systeem is de trainingsdata voor de output dan ook videomateriaal. We analyseren bij deze videobestanden elke frame, op basis van hun RGB waarden. Vervolgens vertellen we het systeem dat de waarden van de beelden horen bij de geluidsanalyse. We gebruiken in dit systeem twee verschillende neurale netwerken: een volledig verbonden neuraal netwerk en een recurrent neuraal netwerk. Een volledig verbonden neuraal netwerk (zie fig. 6) is een type neuraal netwerk waarbij alle neuronen in elke laag met elkaar verbonden zijn. In zo'n netwerk wordt elke invoerlaag verbonden met elke neuron in de volgende laag, wat resulteert in een hoog aantal verbindingen en parameters. We gebruiken dit netwerk in het audio-naar-video systeem voor de vertaling van de geluidsdata naar beelddata. Een recurrent neuraal netwerk (zie fig. 7) is een type neuraal netwerk waarbij de output van een eerdere stap wordt teruggevoerd naar de input van de huidige stap. Hierdoor kunnen deze netwerken informatie opslaan over



Figuur 6: *Voorbeeld van een volledig verbonden neuraal netwerk, in dit geval met 4 inputs, 1 hidden layer en 5 outputs.*



Figuur 7: *Voorbeeld van een recurrent neuraal netwerk.*

eerdere invoer en deze informatie gebruiken om toekomstige voorspellingen te verbeteren. Dit netwerk gebruiken we in dit geval om sequenties van beelden te maken. De afbeeldingen worden gemaakt op basis van de geluidsanalyse, maar om hier vervolgens een bewegend beeld van te maken zullen er meer frames gegenereerd moeten worden die passen bij de context van de huidige frame, zodat de opeenvolging aan frames niet willekeurig lijkt.

Toekomstvisie

Zoals met veel ontwikkelingen binnen AI, zullen instrumenten die geluid-naar-beeldsynthese mogelijk maken exponentieel beter worden. Dit betekent dat het ook een proces is dat steeds toegankelijker zal worden. Het proces kan nu nog erg traag werken, helemaal in de webbrowser. Dat zal, naar vermoeden, al snel anders worden. Systemen die in staat zijn geluid-naar-beeldsynthese te verrichten zullen in de toekomst zeer waarschijnlijk voor iedereen in realtime beschikbaar zijn in de browser, door slechts naar een site te navigeren. Hierdoor zul je deze techniek ook meer toegepast zien worden in de kunstwereld. De vraag is of het hierdoor een herkenbare truc zal zijn die gauw haar aantrekkingskracht verliest, zoals we al eerder zagen met andere ontwikkelingen binnen AI.

Wanneer geluid-naar-beeldsynthese verder ontwikkelt, zal de esthetiek van de werken die er mee gemaakt worden ook veranderen. Veel kunstenaars die in het verleden gebruik maakten van AI om beeld te genereren deelden dezelfde esthetiek met elkaar. Met elke ontwikkeling van AI lijkt de kunstwereld deze te omarmen en tot in de dood uit te putten. Hierbij kun je denken aan de, nu erg gedateerde, DeepDream afbeeldingen die erg populair waren rond 2017 (zie fig.8). In 2023 zie je dat veel generatieve AI kunst juist hyperrealistisch en gedetailleerd is, omdat dit pas kort geleden mogelijk is geworden (zie fig. 9). Ontwikkelingen op technologische vlak impliceren nieuwe creatieve middelen, toch is het belangrijk om het middel te scheiden van het doel. Het middel bepaalt niet de esthetiek van het werk, de kunstenaar bepaalt de esthetiek met de hulp van het middel.

Conclusie

AI maakt veel complexe taken die voorheen onhaalbaar leken mogelijk. Audio vertalen naar beeld is hier een goed voorbeeld van, al zul je zelf een vertaalmethode moeten vinden die qua esthetiek interessante resultaten oplevert. Geluid-naar-beeldsynthese is een techniek die als kunstenaar veel creatieve opties kan bieden.

Door het leggen van een connectie tussen twee modaliteiten die niet in direct verband staan met elkaar maak je een systeem waarvan het resultaat enorm kan verrassen. Met geluid-naar-beeldsynthese komen namelijk verschillende disciplines samen. Het is nu

Figuur 8: *Voorbeeld van DeepDream (bron: TensorFlow).*

Figuur 9: *Werk uit 'The Lumina Creature Collection' (2023), van Krista Kim.*

mogelijk om als geluidsontwerper schilderijen te maken met je werk. Muziek kan gecreëerd worden voor haar visuele esthetiek, in plaats van haar auditieve karakteristieken. Het is mogelijk deze techniek in je website te integreren, maar de problemen van de limitaties van de webbrowser zul je creatief moeten oplossen. Het gebruiken van geluid-naar-beeldsynthese is op dit moment nog interessant, omdat het nog een erg nieuw fenomeen is, maar de uitdaging ligt bij de kunstenaar om met deze techniek interessant werk te maken.

Bronvermelding

1. He Huang, Philip S. Yu and Changhu Wang "An Introduction to Image Synthesis with Generative Adversarial Nets"(2018)
2. Chia-Hung Wan, Shun-Po Chuang, Hung-Yi Lee "Towards Audio to Scene Image Synthesis using Generative Adversarial Network."(2018)
3. Xinsheng Wang, Tingting Qiao, Jihua Zhu1, Alan Hanjalic2, Odette Scharenborg "S2IGAN: Speech-to-Image Generation via Adversarial Learning"(2020)
4. TensorFlow "DeepDream"(2022)
5. Marije Baalman "Composing Interactions, An Artist's guide to Building Expressive Interactive Systems"(2022)
6. Hugh Rawlinson, Nevo Segal, Jakub Fiala "Meyda: an audio feature extraction library for the Web Audio API"(2014)