One-shot learning van gebaren in een convolutioneel neuraal netwerk

Jasper Vaneessen

Promotor: prof. dr. ir. Joni Dambre
Begeleider: Lionel Pigou

Masterproef ingediend tot het behalen van de academische graad van
Master of Science in de industriële wetenschappen: informatica

Vakgroep Elektronica en Informatiesystemen
Voorzitter: prof. dr. ir. Rik Van de Walle
Faculteit Ingenieurswetenschappen en Architectuur
Academiejaar 2016-2017

UNIVERSITEIT
GENT

Toelating tot bruikleen

Jasper Vaneessen, juni 2017

Voorwoord

Jasper Vaneessen, juni 2017

Overzicht

One-shot learning van gebaren in een convolutioneel neuraal netwerk

Jasper Vaneessen

Promotor: Prof. Dr. Ir. Joni Dambre
Begeleider: Ir. Lionel Pigou

Masterproef ingediend tot het behalen van de academische graad van
Master of Science in de industriële wetenschappen: informatica

Vakgroep Elektronica en Informatiesystemen
Voorzitter: Prof. Dr. Ir. Rik Van de Walle

Faculteit Ingenieurswetenschappen & Architectuur
Academiejaar 2016-2017 Universiteit Gent

Trefwoorden

gebarenherkenning, convolutionele neurale netwerken, machine learning, deep learning, one-shot learning

Inhoudsopgave

Afkortingen

ANN	Artificiëel Neuraal Netwerk
CNN	Convolutioneel Neuraal Netwerk

Hoofdstuk 1

Inleiding

1.1 Gebarentaal

Gebarentaal is in de eerste plaats een taal. Taal is een begrip dat moeilijk te definiëren valt en de meeste pogingen hiertoe beperken zich tot gesproken taal. Een definitie die ook voor gebarentaal kan gebruikt worden is: een natuurlijk ontstaan communicatiemidddel waarmee je kan communiceren over alles wat je denkt, ziet, voelt en droomt. [**?**]

Gesproken taal en gebarentaal verschillen in de manier waarop gecommuniceerd wordt: oraal-auditief tegenover gestueel-visueel. Door middel van hand-, hoofd- en armbewegigingen wordt een woord "uitgesproken" en vervolgens visueel waargenomen.

Een gebarentaal onstaat, net zoals een gesproken taal, spontaan en natuurlijk door contact tussen mensen. Net door deze spontane ontwikkeling is er geen universele gebarentaal. Evenals we verschillende gesproken talen en dialecten kennen per land of regio zijn er ook verschillende gebarentalen [Van Herreweghe en Vermeerbergen, 2009]. In Nederland is er bijvoorbeeld de Nederlandse Gebarentaal (NGT) en in België de Vlaamse Gebarentaal (VGT) en de Waalse Gebarentaal (la Langue des Signes de Belgique Francophone, LSFB). VGT verschilt dan weer van provincie tot provincie, met de grootste verschillen tussen West-Vlaanderen en Limburg, de twee verst uiteenliggende regio's.

Een gebarentaal heeft een eigen grammatica en lexicon. Het lexicon of de gebarenschat is de verzameling van alle woorden of gebaren in de taal. Het lokale gebarenschat moet volledig onafhankelijk van het lokale woordenschat worden beschouwd.
Bepaalde woorden uit de ene taal kunnen niet eenduidig vertaald worden in een andere taal. Het woord "gezelligheid"kent bijvoorbeeld geen Engelse vertaling en voor het Duitse "fingerspitzengefühl"hebben we in de Nederlandse taal ook geen alternatief.
Tussen een gebarentaal en een gesproken taal geldt dezelfde verhouding. Er is niet altijd een een-op-een relatie tussen een woord en een gebaar.

Communicatie tussen doven en horenden is vaak een struikelblok. Sommige doven kunnen liplezen en zo opmaken wat een spreker wil vertellen. Voorwaarde hierbij is dat de spreker goed moet articuleren en natuurlijk niet te snel spreekt.
Er kan ook altijd schriftelijk gecommunciceerd worden maar dit is een erg trage en onpersoonlijke vorm van communicatie. Ook is de bedrevenheid van een dove persoon in het schrijven van een gesproken taal vaak lager dan die van een horende.
Doven kunnen zich ook beroepen op een tolk. Dit kan een vriend zijn die horende is en gebarentaal kent of een beroepstolk. In Vlaanderen kunnen doven terecht bij het Vlaams Communicatie Assistentie Bureau voor Doven (CAB) om een tolk in te huren. De Vlaamse overheid betaalt een aantal tolkuren terug. Onder andere achttien tolkuren voor privédoeleinden, achttien voor sollicitaties en een situatie-afhankelijk aantal tolkuren voor arbeid en beroepsopleiding.

1.2 Automatische gebarentaalherkenning

Er is een communicatieprobleem tussen doven en horenden omdat ze niet dezelfde taal spreken. Er zijn ook vele verschillende gebarentalen en dialecten waardoor er tussen doven onderling ook niet altijd vlot gecommuniceerd wordt. Door het gebruik van hedendaagse technologie moet het mogelijk zijn hierin te helpen en een automatisch herkenningssysteem uit te werken waarmee gebaren in real-time kunnen vertaald worden.

Het herkennen van objecten of gebaren is iets waar de mens niet bij stilstaat. Een pasgeboren kind begint vanaf het openen van de ogen zijn waarneming en herkenningsvermogen te trainen. Terwijl we leren organiseren we vormen, objecten en categoriën in nuttige taxonomiën en linken deze dan later naar onze taal [Fei-Fei et al., 2006]. Eenmaal de leeftijd van zes jaar bereikt is kan een kind bijvoorbeeld 104 objectcategoriën onderscheiden zonder hierbij stil te staan.

Als mens kunnen we gebaren makkelijk differentiëren door registratie van armbewegingen, mimiek, houding van de handen en de manier waarop vingers gestrekt of geplooid worden. De neurologische fenomenen die deze vaardigheden kunnen verklaren worden nog steeds onderzocht.

Een machine of computer kan zien via het gebruik van een camera. Een beeld wordt voorgesteld door een matrix met pixelwaarden die de lichtintensiteit op dat bepaalde punt weergeeft. Traditioneel zijn er grijswaarden- en kleurbeelden maar tegenwoordig wordt ook vaak gebruikt gemaakt van 3D-cameratechnologiën, zoals de Microsoft Kinect [Kühn, 2011], zodat er een aanvullend dieptebeeld is. Deze beelden gelden dan als de visuele data voor het systeem, daarna moeten specifieke technologiÃ«n worden ingezet om nuttige informatie uit deze data te halen.

Een automatisch herkenningssysteem zal moeten leren omgaan met de grote variabiliteit van de invoer. De gebaren die het moet herkennen zullen uitgevoerd worden door mensen van verschillende groottes. De vlotheid van het gebaren tussen ervaren en beginnende gebarentaligen zal ook sterk verschillen. De persoon zal ook niet altijd mooi recht in het midden van het beeld staan of even ver van de camera. Ook links- en rechtshandigheid heeft een invloed op het gebaren evenals de expressiviteit van de spreker.
De aanwezigheid van andere mensen of veel beweging in de achtergrond bemoeilijkt ook het herkennen van gebaren. Daarboven moet ook nog rekening gehouden worden met de lokale belichting, de spreker kan onderbelicht of overbelicht zijn waardoor bepaalde contouren moeilijker te detecteren vallen.

Een compleet gebarentaalherkenningssysteem zal moeten voorzien in gebarensegmentatie, gebarenherkenning en grammaticale samenstelling van gebaren.

1.2.1 Gebarensegmentatie

Wanneer we een persoon die gebarentaal spreekt registreren met een camera krijgen we een continue stroom aan informatie. In een bepaalde tijdspanne kan een persoon een of meerdere gebaren uitvoeren en het is onbekend waneer een gebaar begint of eindigt. De segmentatie van deze gebaren is dus een eerste uitdaging voor een herkenningssysteem. Er is minder belangstelling naar deze "continue" gebaarherkenningssystemen omdat vaak wordt uitgegaan van vooraf gesegmenteerde beelden [Kuehne et al., 2011].
Tussen elk gebaar zit er een beweging die de overgang vormt tussen twee gebaren: de bewegingsepenthesis. Armen en handen gaan van eindpositie van het eerste gebaar naar beginpositie van het volgende. [Yang et al., 2010] Deze beweging moet gedetecteerd en gefilterd worden willen we een foutloze segmentatie krijgen.

1.2.2 Gebarenherkenning

Eenmaal we weten wanneer een gebaar begint en eindigt kunnen we het gaan identificeren. Uit de verzamelde visuele data wordt nuttige informatie geëxtraheerd waarmee het model kan beslissen over welk gebaar het gaat. Het beeld wordt omgezet in een beeldrepresentatie, bestaande uit een of meerdere featurevectoren. Deze representatie wordt vervolgens gebruikt door een classificatie methode die het een klasselable geeft.

[Guo en Yang, 2017] stelt een gebaarherkenningssyteem voor die zich focust op de handen. Uit het dieptebeeld van een Kinectcamera wordt de hand gesegmenteerd via thresholding. Drie features worden vervolgens bijgehouden en gebruikt: de verandering van de handvorm, de beweging van de hand in het tweedimensionaal vlak en de beweging van de hand in de diepte (z-as).
Er wordt gebruik gemaakt van twee classificatie methodes: Hidden Markov Modellen

(HMM) en Fuzzy Neurale Netwerken (FNN). HMM is een classificatietechniek die rekening houdt met het tijdsaspect. FNN is een combinatie van fuzzy (vage) theorie en artificiële neurale netwerken.

1.2.3 Grammaticale samenstelling

1.3 One-shot learning

[Fei-Fei et al., 2006]

Een kind van zes jaar is al in staat om 104 objectcategoriën te onderscheiden [Biederman, 1987].

1.3.1 Uitbreidbaarheid herkenningssysteem

Een taal is voortdurend in verandering. Het lexicon van een gebarentaal groeit mee met de tijd. Gloednieuwe termen of zaken die voordien geen beschrijving kenden in een gebarentaal worden toegevoegd. Hogescholen en universiteiten gebruikten lange tijd geen gebarentaal waardoor er weinig wetenschappelijke termen opgenomen zijn in de gebarenschat. Gelukkig komen er vandaag steeds meer wetenschappelijke gebaren bij.

Een automatische herkenningssysteem zal moeten leren omgaan met dit groeiende lexicon. Een strategie kan zijn om na verloop van tijd (vanaf een bepaald aantal nieuwe gebaren) het systeem te hertrainen met voorbeelden van de oude gebaren en de nieuwe gebaren. Hierbij wordt er dus vanaf nul gestart en een nieuw model opgebouwd.

Een eerste probleem is het verzamelen van de data. Deep learning methodieken hebben een complexe structuur en erg veel parameters. Om een grote hoeveelheid parameters te optimaliseren voor een taak heb je een grote hoeveelheid data nodig om uit te leren. Als we dus een nieuw gebaar willen bijleren aan een herkenningssysteem hebben we vele voorbeelden nodig van dit ene gebaar, liefst tegen verschillende achtergronden, uitgevoerd door verschillende personen en in verschillende lichtomstandigheden.
Het maken van dergelijke datasets is een erg kostelijke en tijdrovende opdracht.

Het model vanaf nul terug hertrainen vraagt veel tijd en rekenvermogen. Alle vooraf opgedane kennis wordt gewist dus alle tijd en moeite die eerder geïnvesteerd werd is voor niets. Het systeem zal ook minstens evenveel rekentijd nodig hebben als tijdens de opbouw van het vorige model.

Als we zo een aantal keer het herkenningssysteem willen uitbreiden zullen we veel kostbare tijd en energie verspillen.

1.3.2 Bijleren bij mensen

1.3.3 One-shot learning in de literatuur

[Lake et al., 2011] stelt een generatief model voor voor het herkennen van handgeschreven karakters. Het vertrekt vanuit de notie dat de mens een teken schrijft in verschillende halen of lijnen en ook zo een nieuw teken leert herkennen. Er wordt een dataset opgebouwd van 1600 tekens die door verschillende gebruikers online geregistreerd worden. Elke lijn die een gebruiker plaatst wordt opgeslaan alsook de volgorde van tekenen. Zo bestaat elk teken uit een opeenvolging van lijnen met verschillende vorm en lengte. De verzameling van al deze lijnen wordt gebruikt als voorafgaande kennis om nieuwe tekens bij te leren met een voorbeeld. Het nieuwe teken wordt door het systeem opgedeeld in lijncomponenten die dan afgetoetst worden tegen het model. Zo ontstaat een nieuwe representatie voor het bijgeleerde gebaar die kan gebruikt worden voor herkenning. Er wordt een nauwkeurigheid van 54.9 % behaald tegenover 39.6 % voor een implementatie aan de hand van Deep Boltzman Machines (DBM). Wanneer bij het aanleren van het nieuwe gebaar de lijninformatie van de dataset wordt gebruikt in plaats van die van het systeem zelf wordt een nauwkeurigheid van 63.7 % waargenomen.

[Wu et al., 2012] buigt zich over de ChaLearn One-shot Learning Gesture Challenge 2011 en leert vanuit slechts een voorbeeld een gebaar te herkennen zonder enige voorgaande kennis. Er wordt geëxperimenteerd met een aantal feature descriptors en classificatie metho-

des waaruit Extended Motion History Images (Extended MHI) en Maximum Correlation Coëfficiënt (MCC) als best presterende worden gevonden. Extended MHI bestaat zelf uit drie representaties: MHI en Inversed recording (INV) focussen zich op bewegingsinformatie respectievelijk in het begin en op het einde van het gebaar terwijl Gait Energy Information (GEI) repetitieve beweging registreert. Het systeem behaalt een Levensteihnafstand van 0.29685 (tussen O en 1 waarbij 0 optimaal) op de validatieset en presteert zeer goed op gebaren waarin er veel beweging is. De twee meer statische gebaren uit de dataset worden het minst goed gededecteerd met een nauwkeurigheid lager dan 45 %.

[Caelles et al., 2016] stelt een CNN voor die uit een voorbeeld de voorgrond van de achtergrond onderscheidt in een video. Het CNN wordt vooraf getraind op de ImageNet dataset. Een dataset van 1,2 miljoen afbeeldingen uit meer dan duizend categoriën. Door deze pre-training op een zeer ruime dataset is het model algemeen en leert het eigenlijk wat 'een object' is. Hierna wordt het model verfijnd voor het volgen van een voorgrondsobject uit een video. Het eerste frame van de video wordt gemaskeerd en hierop stelt het model zich af. Deze architectuur verbetert de state-of-the-art op de Densely Annotated Video Segmentation (DAVIS) dataset met 11.2 % (79.8% vs 68.0%).

1.4 Doelstelling

Hoofdstuk 2

Technische aspecten

2.1 Machine Learning

Leren is een veelzijdig fenomeen dat bestaat uit verschillende processen: het verkrijgen van declaratieve kennis, het ontwikkelen van motorische en cognitieve vaardigheden door instructie en ervaring, het organiseren van nieuwe kennis in algemene representaties en het ontdekken van nieuwe feiten via observatie en experimentatie.

Sinds het begin van het computertijdperk proberen onderzoekers het menselijk leren na te bootsen en deze processen te vertalen naar de informatietheorie. Het machinaal leren is nog steeds een erg uitdagend doel in de kunstmatige intelligentie (KI).

We kunnen zeggen dat een computerprogramma of machine leert als het zijn performantie op een bepaalde taak verbetert met ervaring [Carbonell et al., 1983]

In het geval van dit onderzoek is deze taak het classificeren van gebaren. Het model zal dus leren

2.2 Artificiëel neuraal netwerk

2.3 Convolutioneel neuraal netwerk

2.4 Hyperparameters

Bibliografie

[Ba et al., 2015] Ba, J., Swersky, K., Fidler, S., en Salakhutdinov, R. (2015). Predicting Deep Zero-Shot Convolutional Neural Networks using Textual Descriptions. *arXiv:1506.00511 [cs]*. arXiv: 1506.00511.

[Biederman, 1987] Biederman, I. (1987). Recognition-by-Components: A Theory of Human Image Understanding. *Psychological Review*, 94(2):115–147.

[Braun et al., 2009] Braun, D. A., Aertsen, A., Wolpert, D. M., en Mehring, C. (2009). Motor Task Variation Induces Structural Learning. *Current Biology*, 19(4):352–357.

[Caelles et al., 2016] Caelles, S., Maninis, K.-K., Pont-Tuset, J., Leal-TaixÃ©, L., Cremers, D., en Van Gool, L. (2016). One-Shot Video Object Segmentation. *arXiv:1611.05198 [cs]*. arXiv: 1611.05198.

[Carbonell et al., 1983] Carbonell, J. G., Michalski, R. S., en Mitchell, T. M. (1983). An Overview of Machine Learning. In Michalski, R. S., Carbonell, J. G., en Mitchell, T. M., editors, *Machine Learning*, Symbolic Computation, pages 3–23. Springer Berlin Heidelberg. DOI: 10.1007/978-3-662-12405-5_1.

[Ciresan et al., 2012] Ciresan, D. C., Meier, U., en Schmidhuber, J. (2012). Transfer learning for Latin and Chinese characters with deep neural networks. In *The 2012 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6. IEEE.

[Cooper et al., 2012] Cooper, H., Ong, E.-J., Pugeault, N., en Bowden, R. (2012). Sign language recognition using sub-units. *Journal of Machine Learning Research*, 13(Jul):2205–2231.

[Fei-Fei et al., 2006] Fei-Fei, L., Fergus, R., en Perona, P. (2006). One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594–611.

[Guo en Yang, 2017] Guo, X.-L. en Yang, T.-T. (2017). Gesture recognition based on HMM-FNN model using a Kinect. *Journal on Multimodal User Interfaces*, 11(1):1–7.

[Hoffman et al., 2013] Hoffman, J., Tzeng, E., Donahue, J., Jia, Y., Saenko, K., en Darrell, T. (2013). One-Shot Adaptation of Supervised Deep Convolutional Models. *arXiv:1312.6204 [cs]*. arXiv: 1312.6204.

[Ji et al., 2013] Ji, S., Xu, W., Yang, M., en Yu, K. (2013). 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231.

[Kadrev et al., 2016] Kadrev, G., Kostadinov, G., en Ruskov, P. (2016). Expansion of a CNN-based image classifier's scope using transfer learning and k-NN. Technical report. In *2016 IEEE 8th International Conference on Intelligent Systems (IS)*, pages 764–770.

[Karpathy et al., 2014] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., en Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732.

[Kuehne et al., 2011] Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., en Serre, T. (2011). Hmdb: a large video database for human motion recognition. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2556–2563. IEEE.

[Kühn, 2011] Kühn, T. (2011). The kinect sensor platform. *Proc. Advances in Media Technology*, pages 1–4.

[Lake et al., 2011] Lake, B. M., Salakhutdinov, R., Gross, J., en Tenenbaum, J. B. (2011). One shot learning of simple visual concepts. In *CogSci*, volume 172, page 2.

[Lake et al., 2015] Lake, B. M., Salakhutdinov, R., en Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338.

[Lima et al., 2017] Lima, E., Sun, X., Dong, J., Wang, H., Yang, Y., en Liu, L. (2017). Learning and Transferring Convolutional Neural Network Knowledge to Ocean Front Recognition. *IEEE Geoscience and Remote Sensing Letters*, 14(3):354–358.

[Michalski et al., 2013] Michalski, R. S., Carbonell, J. G., en Mitchell, T. M. (2013). *Machine Learning: An Artificial Intelligence Approach*. Springer Science & Business Media.

[Oquab et al., 2014] Oquab, M., Bottou, L., Laptev, I., en Sivic, J. (2014). Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1717–1724.

[Pan en Yang, 2010] Pan, S. J. en Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

[Pigou,] Pigou, L. Gebarentaalherkenning met convolutionele neurale.

[Samuel, 1959] Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, 3(3):210–229.

[Santoro et al., 2016] Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., en Lillicrap, T. (2016). One-shot Learning with Memory-Augmented Neural Networks. *arXiv preprint arXiv:1605.06065*.

[Van Herreweghe en Vermeerbergen, 2009] Van Herreweghe, M. en Vermeerbergen, M. (2009). Flemish Sign Language standardisation. *Current Issues in Language Planning*, 10(3):308–326.

[Woodward en Finn, 2017] Woodward, M. en Finn, C. (2017). Active One-shot Learning. *arXiv:1702.06559 [cs]*. arXiv: 1702.06559.

[Wu et al., 2016] Wu, D., Pigou, L., Kindermans, P.-J., Nam, L. E., Shao, L., Dambre, J., en Odobez, J.-M. (2016). Deep Dynamic Neural Networks for Multimodal Gesture Segmentation and Recognition.

[Wu en Shao, 2014] Wu, D. en Shao, L. (2014). Deep dynamic neural networks for gesture segmentation and recognition. In *Workshop at the European Conference on Computer Vision*, pages 552–571. Springer.

[Wu et al., 2012] Wu, D., Zhu, F., en Shao, L. (2012). One shot learning gesture recognition from RGBD images. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 7–12.

[Yang et al., 2010] Yang, R., Sarkar, S., en Loeding, B. (2010). Handling Movement Epenthesis and Hand Segmentation Ambiguities in Continuous Sign Language Recognition Using Nested Dynamic Programming. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(3):462–477.

[Zaki en Shaheen, 2011] Zaki, M. M. en Shaheen, S. I. (2011). Sign language recognition using a combination of new vision based features. *Pattern Recognition Letters*, 32(4):572–577.

[Zeiler en Fergus, 2014] Zeiler, M. D. en Fergus, R. (2014). Visualizing and Understanding Convolutional Networks. In Fleet, D., Pajdla, T., Schiele, B., en Tuytelaars, T., editors, *Computer Vision - ECCV 2014*, Lecture Notes in Computer Science, pages 818–833. Springer International Publishing. DOI: 10.1007/978-3-319-10590-1_53.