ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Exploring the Importance of Sign Language Phonology for a Deep Neural Network

Javier Martínez Rodríguez, Martha Larson, and Louis ten Bosch

Center for Language Studies
Radboud University, Netherlands

Abstract. We conduct an initial investigation to gain insight into whether a deep neural network learns phonological aspects of sign language when classifying video recordings of isolated signs from a continuous signing scenario. We train a series of neural networks to distinguish pairs of signs in Dutch Sign Language, controlling the phonological difference between the signs in each pair. Our results suggest that the intrinsic dimension of the final hidden layer of a network is surprisingly insensitive to the phonological difference between the signs in a pair. However, the ability of the network to discriminate two signs shows a clear trend towards increasing with increasing phonological distinctiveness.

1 Introduction

In spoken language, phonology describes the basic units of sound that contribute to the meaning of words. In parallel, in sign language, phonology describes basic aspects of form. Two signs build a minimal pair if they differ in a single aspect of form and if the difference is large enough to give them two different meanings. In this paper, we conduct an initial investigation to understand the extent to which a deep neural network captures phonological aspects of sign language, i.e., whether the network is learning the same basic units of form that linguists use to describe the fundamental phonological structure of sign language.

Our contribution consists of the results of two exploratory experiments. The first (Section 4.1) suggests that existing measures of intrinsic dimension of neural networks are not correlated with phonological structure. The second (Section 4.2) shows that the ability of the network to discriminate between two signs is correlated with increasing phonological distance. Our paper provides first evidence that deep neural networks are sensitive to phonology, i.e., to the same differences that people use when interpreting the meaning of a sign. However, our results suggest that the deep neural network is also capturing other aspects as it learns to discriminate signs. More information can be found in the original work on which this paper is based [1] and in its implementation on Github[1].

2 Related Work

Isolated sign language recognition. Sign language recognition (SLR) is the problem of recognizing and identifying a particular sign in a video clip. In this paper, we study isolated SLR, also known as word-level SLR, which

[1] `https://github.com/JavierMartnz/MindTheLinguisticGap`

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

uses pre-segmented video clips that contain a single signing instance. Recent isolated SLR approaches rely on either the use of raw RGB videos or pose-based landmarks as input data. Pose-based approaches [2] are able to abstract away from irrelevant visual information like the clothing of a signer or the background. However, 3D-CNNs, and more concretely I3D models [3] using raw RGB video data, outperform pose-based pipelines [4, 5, 6]. We study the I3D model with the same pixel resolution (224 x 224) that is common in the literature.

Intrinsic dimension estimation. The intrinsic dimension (ID) of a dataset refers to the minimal M-dimensional manifold on which the data lies entirely without information loss. ID estimation is an open problem that aims at finding a lower bound for M. The application of ID estimation to real image data is rare and, to the best of our knowledge, limited to the use of maximum likelihood estimation (MLE) for estimating the ID of data [7] and to the use of TwoNN [8] and DeepMDS [9] algorithms to estimate the ID of neural representations. In our experiments we only use MLE and TwoNN, avoiding the computational overhead of training DeepMDS. We study the ID of neural representations, but unlike [8], we modify the task and not the classifier used. We assume that the ID is lower when a task needs fewer conceptual dimensions to be solved.

3 Experimental Setup

Datasets. In this work, we use data from the Corpus Nederlandse Gebarentaal (CNGT) [10, 11], the first continuous signing, linguistically-motivated [12, 13] Dutch Sign Language (DSL) dataset. In CNGT, native DSL signers converse in studio conditions [11]. This dataset contains 72 hours of continuous sign language videos recorded at 25 fps, featuring a total of 92 signers, of which about 12 hours are annotated with sign glosses from the NGT Signbank [14].

Model architecture. The Two-Stream Inflated 3D Con-vNet (I3D) model is an action recognition model [3] that has been used in different sign language tasks such as recognition [4, 5, 6], segmentation [15], feature extraction [16], and sign spotting [17]. It features a cascade of *inflated* inception modules and pooling operations that allow the spatio-temporal processing of input videos at different scales. This implies that predictions are done at a video-level, instead of averaging predictions at frame-level, which we expect allows to capture the temporal structure of sign language. In our experiments, we study the I3D version that uses only RGB video flow, as done by [6]. We follow previous research [5, 6, 15] in using a model pre-trained on the Kinetics dataset [18] and fine tune on our sign language training data. We fine tune the last hidden layer of the model while keeping the rest of the layers frozen to ensure that all learning of sign language is embedded in the representations of this last layer. After exploration, we determine that the average length of signs in the CNGT is approximately 12 frames, corresponding to previous finds that co-articulated signs have length of approximately 7 to 13 frames [19, 20, 21]. Thus, we set the number of frames per input video to 16.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Data preparation. Gloss annotations are used to extract isolated sign video clips from the video footage. For all the experiments, the data is stratified by signer and split in training, validation, and test sets in a 4:1:1 ratio. Training data is transformed with affine transformations, horizontal flips, and by introducing color jitter. Perspective of the videos is also modified to introduce variability. Resolution is reduced by a factor of 0.875 by means of random cropping for the training data and by centered cropping for validation and test splits.

Phonological distance. We propose a measure of phonological distance based on 14 phonological specifications available in NGT Signbank (more details in [1]). These specifications take a large step of abstraction away from concrete articulatory detail [22]. Note that there is no commonly agreed phoneme inventory for sign language, but consensus exists on the basic phonological aspects. We define a phonological distance ph_{dist} that ranges between 1 and 14, based on the number of phonological differences between two signs.

4 Experiments and Results

4.1 Intrinsic dimension experiment

We compare two neural networks trained to distinguish two pairs of signs to gain insight into the impact of phonology on the ID of the representations learned by the network. We calculate the ID on the last hidden layer, which is also the fine-tuned layer. Recall that we expect information specific to sign language to be represented in this layer. Results are reported in Table 1. The first pair

Gloss pair (#data samples)	ph_{dist}	Configuration	Accuracy	Precision	Recall	F-score	ID_{MLE}	ID_{TwoNN}
[GEBAREN-A, JA-A] (1553, 1133)	7	Original	0.89	0.91	0.89	0.90	25.36	20.42
		512 resolution	0.85	0.86	0.90	0.88	**18.36**	**16.81**
		12 fps	0.82	0.84	0.85	0.84	20.92	17.00
[GAAN-NAAR-A, NU-A] (541, 504)	1	Original	0.85	0.84	0.86	0.85	24.34	19.57
		512 resolution	0.76	0.77	0.72	0.74	21.25	18.91
		12 fps	0.70	0.70	0.66	0.68	**19.27**	**15.23**

Table 1: Performance of the classification task on the test set for different input data and network configurations.

`[GEBAREN-A, JA-A]` has $ph_{dist} = 7$ and was chosen because of its high frequency of occurrence in the data. The second pair `[GAAN-NAAR-A, NU-A]` is the most frequent minimal pair ($ph_{dist} = 1$). Note that glosses consist of one or more capitalized Dutch words, with hyphenated extensions, such as a letter indicating a regional variant. Given our previously-stated assumption on ID behaviour, we expect that the hidden layer would need to capture less phonological information in the case of the minimal pair, which would result in a lower ID. Comparing the line marked 'Original' for both pairs we see that there is no substantial difference in ID, suggesting that either the layer does not capture phonology or that the ID is sensitive to something else. To achieve further insight, we change the model and data configurations in two ways that do not impact the meaning of the sign, i.e., leave the phonology untouched. The line marked '512 resolution' reports

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

results when the data has been spatially upsampled (from 256x256 to 512x512) and the line marked '12 fps' reports results when the training data has been temporally down-sampled (from 25 to 12 fps). We see that this manipulation of the data impacts the ID quite markedly. We conclude that the ID does react to changes in the data, but there is not a clear reaction to phonology. In the next experiment, we move on from ID and focus on connecting phonology directly to the performance of the machine learning classifier.

4.2 Phonological distance experiment

We compare a series of neural networks trained to distinguish different pairs of signs. The pairs are chosen to have a steadily increasing phonological distance from $ph_{dist} = 1$ (minimal pair) to $ph_{dist} = 10$. To ensure that our models are well trained, we focus on cases in which a maximum amount of training data is available. We identify four *reference signs* that can be paired with ten *comparison signs* that have increasing ph_{dist}. The reference signs are given in the first line of Table 2, with the comparison signs listed below. For each reference sign, we train ten binary classifiers: reference sign vs. each of the comparison signs. Both reference and the comparison signs are chosen to maximize the amount of data available for training each classifier (at the expense of having a reduced variety of pairs). For each reference sign, we ensure that all ten classifiers are trained on the same number of samples. We randomly sample the training data for all comparison signs to match the number of samples available for the comparison sign with the lowest number of samples.

ph_{dist}	PT-1hand:1 (377 samples each)	WETEN-A (377 samples each)	DOOF-B (350 samples each)	DOOF-A (234 samples each)
1	WETEN-A	PT-1hand:1	DOOF-A	PT-1hand:1
2	DOOF-B	DOOF-B	PT-1hand:1	CI-A
3	PT-1hand	PT-1hand	OOK-A	PT-1hand
4	HOREN-A	HOREN-A	PT-1hand	ZEGGEN
5	GOED-A	GOED-A	GOED-A	GOED-A
6	JA-A	JA-A	JA-A	JA-A
7	PO	PO	PO	PO
8	KLAAR-A	KLAAR-A	GEHANDICAPT-A	KLAAR-A
9	TWIJFEL-A	TWIJFEL-A	TWIJFEL-A	TWIJFEL-A
10	GEBAREN-A	GEBAREN-A	GEBAREN-A	GEBAREN-A

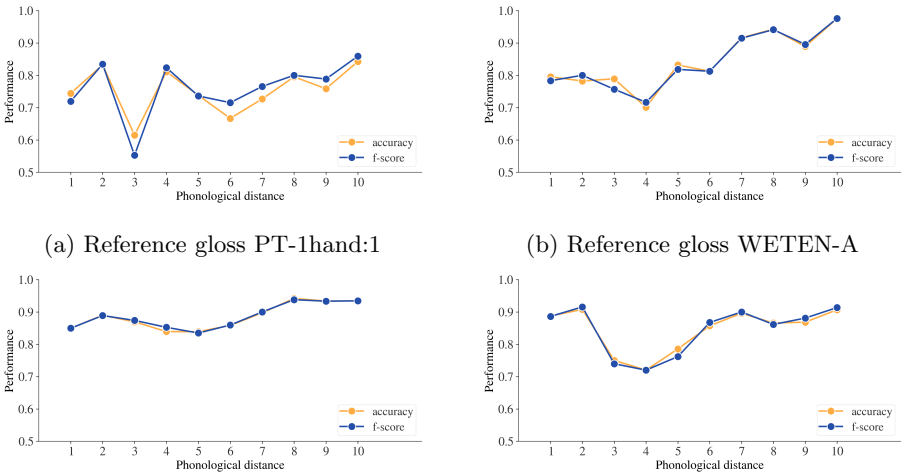
Table 2: Reference signs (top row) and signs with ph_{dist} 1 to 10 (columns).

The performance of the classifiers are shown in Fig. 1. For each reference sign we see an upward trend; namely, performance increases as the phonological distance increases. Exceptions to the trend are discussed in [1] together with additional experiments, not included in Fig. 1, in which we found that ID showed no such correspondence to performance.

5 Conclusion and Outlook

In this paper, we have reported on an initial investigation into the extent to which a deep neural network trained to distinguish pairs of signs learns sign

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.



(a) Reference gloss PT-1hand:1

(b) Reference gloss WETEN-A

(c) Reference gloss DOOF-B

(d) Reference gloss DOOF-A

Fig. 1: Performance on the test set with respect to phonological distance.

language phonology. We have found that intrinsic dimension does not provide direct evidence that phonology is being captured, but that the performance of the neural network becomes better as phonological distance increases. Ideally, deep neural networks would capture phonology, which would help to make them interpretable in the same terms in which people themselves understand meaning in sign language. Future work should further investigate the different mechanisms through which the network is able to learn phonological properties of sign language in order to improve performance and interpretability.

References

[1] Javier Martínez Rodríguez. Mind the Linguistic Gap. Master's thesis, Radboud University, Nijmegen, Netherlands, May 2023. Available at `https://theses.ubn.ru.nl/handle/123456789/15031`.

[2] Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. Skeleton aware multi-modal sign language recognition. In *CCVP*, 2021.

[3] João Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *CVPR*, 2017.

[4] Hamid Reza Vaezi Joze and Oscar Koller. MS-ASL: A Large-Scale Data Set and Benchmark for Understanding American Sign Language. *arXiv preprint arXiv:1812.01053*, 2018.

[5] Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *WACV*, 2020.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

[6] Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues. In *ECCV*, 2020.

[7] Phillip Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. *arXiv preprint arXiv:2104.08894*, 2021.

[8] Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. *NeurIPS*, 2019.

[9] Sixue Gong, Vishnu Naresh Boddeti, and Anil K Jain. On the intrinsic dimensionality of image representations. In *CVPR*, 2019.

[10] Onno Crasborn, Inge Zwitserlood, and Johan Ros. The Corpus NGT. An open access digital corpus of movies with annotations of Sign Language of the Netherlands. Center for Language Studies, Radboud University Nijmegen, 2008.

[11] Onno Crasborn and Inge Zwitserlood. The Corpus NGT: an online corpus for professionals and laymen. *Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages*, pages 44–49, 2008.

[12] Oscar Koller, Jens Forster, and Hermann Ney. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *CVIU*, 2015.

[13] Jens Forster, Christoph Schmidt, Thomas Hoyoux, Oscar Koller, Uwe Zelle, Justus H Piater, and Hermann Ney. RWTH-PHOENIX-weather: A large vocabulary sign language recognition and translation corpus. In *LREC*, 2012.

[14] Onno Crasborn et al. NGT dataset in Global Signbank. Center for Language Studies, Radboud University Nijmegen, 2020.

[15] Katrin Renz, Nicolaj C Stache, Samuel Albanie, and Gül Varol. Sign language segmentation with temporal convolutional networks. In *ICASSP*, 2021.

[16] Tao Jiang, Necati Cihan Camgöz, and Richard Bowden. Looking for the Signs: Identifying Isolated Sign Instances in Continuous Video Footage. In *International Conference on Automatic Face and Gesture Recognition*, 2021.

[17] Ryan Wong, Necati Cihan Camgöz, and Richard Bowden. Hierarchical I3D for Sign Spotting. In *ECCV*, 2023.

[18] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sanja Vijayanarasimhan, Francis Viola, Tom Green, and Tim Back. The kinetics human action video dataset. In *CPVR*, 2017.

[19] Patrick Buehler, Andrew Zisserman, and Mark Everingham. Learning sign language by watching TV (using weakly aligned subtitles). In *CVPR*, 2009.

[20] Tomas Pfister, James Charles, and Andrew Zisserman. Large-scale Learning of Sign Language by Watching TV (Using Co-occurrences). In *BMVC*, 2013.

[21] Ville Viitaniemi, Tommi Jantunen, Leena Savolainen, Matti Karppa, and Jorma Laaksonen. S-pot - a benchmark in spotting signs within continuous signing. In *LREC*, 2014.

[22] Onno Crasborn. *Sign Language: An International Handbook*, volume 37, chapter 2, pages 4–20. De Gruyter Mouton, Pfau, Roland and Steinbach, Markus and Woll, Bencie edition, 2012.