Exploring Latent Sign Language Representations with Isolated Signs, Sentences and In-the-Wild Data

Fredrik Malmberg[1], **Anna Klezovich**[1], **Johanna Mesch**[2], **Jonas Beskow**[1]
[1]Division of Speech, Music and Hearing, KTH, Sweden
[2]Department of Linguistics, Stockholm University, Sweden
fmalmb@kth.se, annkle@kth.se,
johanna.mesch@ling.su.se, beskow@kth.se

Abstract

Unsupervised representation learning offers a promising way of utilising large unannotated sign language resources found on the Internet. In this paper, a representation learning model, VQ-VAE, is trained to learn a codebook of motion primitives from sign language data. For training, we use isolated signs and sentences from a sign language dictionary. Three models are trained: one on isolated signs, one on sentences, and one mixed model. We test these models by comparing how well they are able to reconstruct held-out data from the dictionary, as well as an in-the-wild dataset based on sign language videos from YouTube. These data are characterized by less formal and more expressive signing than the dictionary items. Results show that the isolated sign model yields considerably higher reconstruction loss for the YouTube dataset, while the sentence model performs the best on this data. Further, an analysis of codebook usage reveals that the set of codes used by isolated signs and sentences differ significantly. In order to further understand the different characters of the datasets, we carry out an analysis of the velocity profiles, which reveals that signing data in-the-wild has a much higher average velocity than dictionary signs and sentences. We believe these differences also explain the large differences in reconstruction loss observed.

Keywords: sign language data, VQ-VAE, Representation Learning, Pose Codebook

1. Introduction

Sign languages play a critical role in the communication of deaf communities worldwide, with over 300 different sign languages in use. Despite their significance, sign languages are generally under-resourced compared to spoken languages, with small corpora and limited lexicon due to the need for a manual gloss annotation of sign language videos. While processing of written and spoken languages has advanced rapidly in recent years, with technology performing on par with humans, the same trend has not yet been observed in sign language processing.

Recent progress in speech and text processing has been possible thanks to self-supervised representation learning methods that can be carried out on vast corpora without the need for manual annotation. It has been shown for a speech generation task that learning a powerful data representations significantly improves speech generation (Baevski et al. (2020), van den Oord et al. (2016)). Importantly, this has also made it possible to train models not only on data specifically recorded for the purpose of language technology, but also on in-the-wild data from various Internet sources such as YouTube, which is very beneficial for the low-resourced domain of sign languages.

In this paper, we are investigating how a Vector Quantized Variational Autoencoder (VQ-VAE) representation learning model can learn a codebook of motion primitives from pose-tracked video data. We train this model both on dictionary signs and short sentences, and we investigate how the model's performance generalizes to sign language data from YouTube. Examples of sequences reconstructed from the model can be seen on our project page[1].

In the future perspective this model can be used for producing sign language data representations that can be used as a stepping stone for the sign language generation task.

2. Related Work

Unsupervised representation learning has been found effective in various data domains, for example using masked language for natural language understanding tasks (Devlin et al., 2018) or audio pre-training for speech recognition (Baevski et al., 2020). For generation tasks in the motion domain, different kinds of probabilistic representation learning schemes, such as VQ-VAEs have been successful. For instance, in the co-speech gestures domain, Yazdian et al. (2022) paper focuses on learning representations for motion primitives with the help of denoising autoencoder (DAE) model that encodes poses into simpler representations, and then these representations are fed as sequences into the second model – VQ-DVAE, that learns mo-

[1] www.speech.kth.se/research/vq-sign

Proceedings of the 11th Workshop on the Representation and Processing of Sign Languages, pages 219–224
25 May 2024. © 2024 ELRA Language Resources Association: CC BY-NC 4.0

tion primitives. In the dance generation domain, Siyao et al. (2022) paper uses VQ-VAE as a step for dance generation. The VQ-VAE learns choreographic motion primitives and then an actor-critic GPT model generates dances out of the motions coherently with the music. For more general motion, Jiang et al. (2023) trains a VQ-VAE to create a motion vocabulary that is then used together with a GPT model for several tasks such as Text-to-Motion, Motion Prediction and Motion-to-Text.

Recently, similar representation learning approaches have been applied also to sign language data for sign language understanding task, e.g. a SignBert paper by Zhou et al. (2021) and newer SignBERT+ by Hu et al. (2023). More specifically, a VQ-VAE model has been applied to sign language in Xie et al. (2022), where the main focus is on sign pose sequence generation using a diffusion model. However, in Xie et al. (2022) the authors encode poses frame by frame in the latent space, and as a result they get encoded key points per frame instead of motion primitives capturing a sequence of frames. In our work we use VQ-VAE as a way to learn a codebook of motion primitives for sign language data.

3. Data

3.1. Swedish Sign Language Dictionary

This study uses the Swedish Sign Language (STS) Dictionary Svenskt teckenspråkslexikon (2024), which contains 21 000 entries and 6700 sentence examples. Each dictionary entry includes a video of the sign, phonological information, variants, and example sentences. The Swedish Sign Language Dictionary is also linked to the Swedish Sign Language Corpus through ID-glosses (Mesch et al., 2012; Mesch and Wallin, 2015). It highlights how this focuses on lexical issues, particularly sign lemmatization, and aims to offer a more comprehensive lexical description and understanding of language use in natural conversation settings. The total duration of the dictionary data is 664 minutes, or 1 731 976 frames.

3.2. YouTube Data

For the purposes of testing our representation learning model on the in-the-wild data, we collected data from the YouTube channel "UR Teckenspråk"[2]. Our YouTube dataset contains 17 videos from the Djupdyk playlist with a total duration of 105.6 minutes and a total number of frames 158 406, which is comparable to the size of our test dataset (99 minutes and 229 802 frames respectively).

[2] www.youtube.com/@URTeckensprak

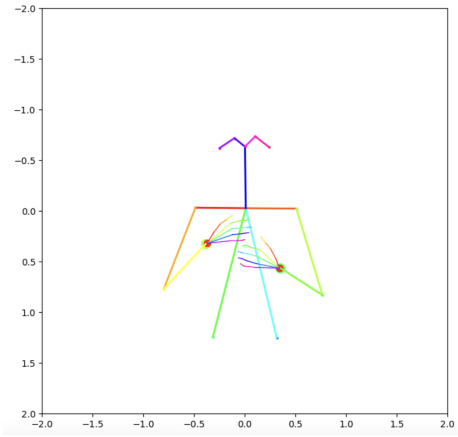

Figure 1: Example of extracted and normalized keypoints. Hands are relocated to wrist positions and lines are drawn between keypoints for illustration purposes.

3.3. Pose Tracking and Pre-Processing

DWPose (Yang et al., 2023) was used to extract 2D pose keypoints frame by frame in the videos. The decision to use DWPose over the commonly used MediaPipe (Zhang et al., 2020) was based on its subjectively perceived robustness for the specific use case. For the sake of simplicity, only keypoints relating to the overall upper body pose, arms and hands were used resulting in 56 2D keypoints per frame (see Figure 1 for an example). In the future perspective, we want to add facial features since most non-manuals are carried out through the facial features.

In order to preprocess the raw pose data, we select the center of the first frame (the keypoint that connects body with the neck) in the sequence in order to shift the bodypose with respect to it, and then we scale the pose by a scaling factor based on the distance between the left and right shoulder. The keypoints related to the hands are shifted so that the wrist is located in the center for each frame in order to capture finger movements and hand shapes regardless of their global position.

3.4. Velocity Profile Examination

Our VQ-VAE model architecture requires choosing the sequence length to encode. Since we wanted to find motion primitives for sign language data, we investigated velocity profiles of the STS dictionary dataset to estimate the appropriate sequence length to encode.

In order to find velocity, we calculated centroids for each hand coordinates and then computed the distance between the centroids of neighboring video frames separately for each hand. Velocity was calculated as an Euclidean norm of the

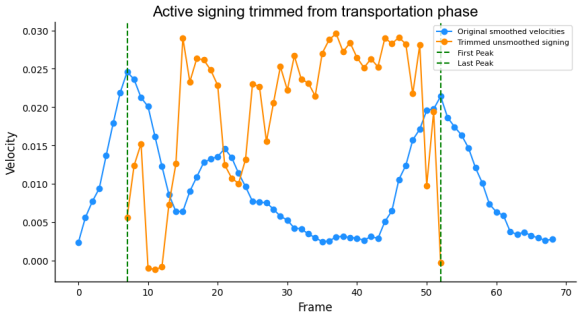

Figure 2: Velocity profile for the STS sign 'kloster'. First peak and last peak signify the transportation phase. Velocity calculated as a distance between hand coordinates centroids for neighboring frames.

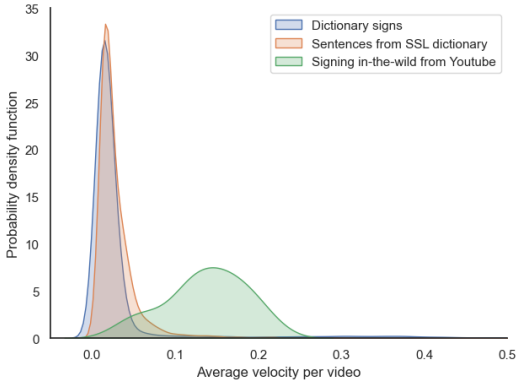

Figure 3: Comparison of average velocity distributions for three types of data.

displacement vector between the centroids and averaged between hands, because we have both left handed and right handed signers in the dataset.

By studying the the number of velocity peaks between preparation for signing and retraction movements, we found that an average number of frames that correspond to one motion in both dictionary signs and sentences is around 30 frames. Similarly to Börstell (2023) we used the moving average to smooth the signal and extract the first and last peaks that correspond to transportation movements (preparation and retraction). For the analysis, we trimmed a signal from transportation movements, and then extracted the peaks from the inverted active signing signal based on a heuristic where the peak is significant if it is higher than one standard deviation from the mean (see example of a velocity profile for sign 'kloster' in Figure 2). As a result, we estimated that the average number of frames for motions in a sentence is 31.7 frames and 26.3 in dictionary signs.

This information was used then in the model design stage, where we assigned sequence length in the VQ-VAE to 30 frames for both signs and phrases based models.

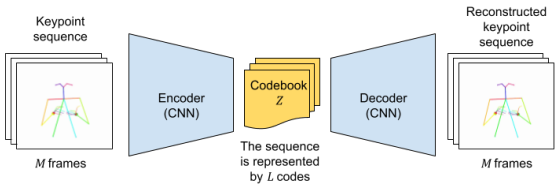

Figure 4: The VQ-VAE consists of an encoder that takes sequences of poses as inputs, a codebook that captures the motion codes and a decoder that outputs reconstructed sequences.

Additionally, we compared the distributions of average velocities for videos in three datasets that we are using (see Figure 3). As a result, we discovered that signing in the wild is much faster than both dictionary sentences and signs, as expected. While the velocity of signing in sentences from a dictionary is only a little bit higher than the velocity of dictionary signs. We expected the velocity of sentences to be closer to the signing-in-the-wild.

4. Model

4.1. VQ-VAE Architecture

Inspired by the architecture in Jiang et al. (2023), that focuses on tokenizing body motion, we train a 2D sign motion tokenizer using a VQ-VAE (van den Oord et al., 2018). It consists of an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$, with a discrete latent representation transforming motion into a structured codebook. $\mathcal{E}$ generates dense motion tokens through a network consisting of 1D convolutions that are quantized using codes from the codebook, which $\mathcal{D}$, also based on 1D convolutions, reconstructs into sequences (see Figure 4). In our model, the encoder takes a sequence of sign poses represented as normalized 2D keypoint coordinates, of length M, and produces latent vectors $\hat{z}^{1:L} = \mathcal{E}\left(m^{1:M}\right)$, effectively capturing sequences of frames in each latent vector and downsampling a motion sequence $L = M/l$, where l is the downsampling factor. These vectors are then discretized into a set of codebook entries z through quantization so that each entry z_i belongs to a learnable codebook $Z = \left\{z^i\right\}_{i=1}^{K} \subset R^d$, with K latent embedding vectors of dimension d.

$$z_i = Q\left(\hat{z}^i\right) := \arg\min_{z_k \in Z} \left\|\hat{z}_i - z_k\right\|_2 . \quad (1)$$

To reconstruct the sequence the decoder uses these embeddings and outputs a sequence of sign poses of length M.

Optimization employs reconstruction loss ($\mathcal{L}_r$), which compares the mean squared error between the input and the output of the VQ-VAE, and a commitment loss ($\mathcal{L}_c$) that ensures the encoder com-

mits to an embedding and limits the growth of the embedding space. We also employ other additional techniques for quality enhancement such as replacing the embedding loss ($\mathcal{L}_e$) that minimizes the difference between the encoded sequences and the closest code embeddings, with exponential moving average (EMA) as in Razavi et al. (2019).

5. Experiments

For the following experiments we used a codebook size, K, of 512 and also set the dimension of the embedding vectors, d, to 512, following Jiang et al. (2023). Based on our analysis of velocity profiles, we used a sequence length, M, of 30 frames, and for the encoder network, we used a depth of 3 and stride 2 resulting in a downsampling factor, l, of 7.5 and a latent encoding of length, L, of 4. This was to ensure that each token in our codebook would correspond to a motion and not only keyframes as in other works such as Xie et al. (2022).

We trained three models on the Swedish Sign Language Dictionary: one using only individual signs (signs model), one using only sentence data (sentences model) and one using all data (mixed model). The data was split 80/10/10 in a train, validation and test set and the same split was used for all models to prevent information leakage.

	Training Dataset		
Test Dataset	Signs	Sentences	Mixed
Signs	0.0067	0.0214	0.0077
Sentences	0.0074	0.0044	0.0074
YouTube	0.0211	0.0146	0.0157

Table 1: Reconstruction loss for models trained on different subsets of the Swedish sign data measured as the mean squared error between the input and the output of the VQ-VAE

As can be seen in Table 1 the models trained on only signs or sentences exhibited better reconstruction for the type of data they were trained on, which was expected. It can also be seen that the reconstruction loss on data from YouTube was lower for the model trained on sentences.

To further investigate how the models learn to represent motion primitives in the codebook, we evaluated the use of codes for the model trained on all the data for 5000 test and training samples from signs and 5000 test samples from sentences respectively. Figure 5 shows that there is a difference in the usage of the codebook and that the distribution over codes for the samples is more similar between signs than between signs and sentences.

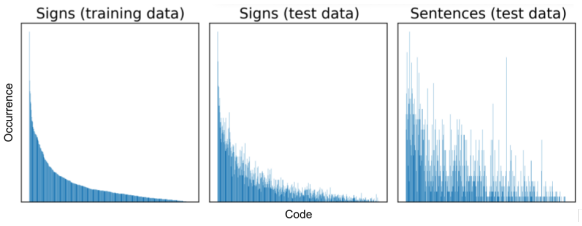

Figure 5: Codebook usage for the model trained on both signs and sentences. The three histograms are sorted horizontally by most used codes for the training data for individual Signs (left).

6. Discussion and Conclusion

The results in this paper indicate that it is possible to capture some of the dynamic nature of signing using an unsupervised model such as a VQ-VAE. As is seen in the reconstruction results between the different models and on the different types of data, it is clear that capturing motion primitives more similar to the dynamic of the target data yields better results (see Table 1).

6.1. Time Dependence

In its current setup, the VQ-VAE model architecture puts fixed limits on a sequence length, which means data is cut and/or padded to deal with different lengths of motions. The previous works usually set a fixed sequence length based on the domain of the data. For instance, the authors of Siyao et al. (2022) use longer sequence length in their model – 240 frames, compared to co-speech gestures paper Yazdian et al. (2022), who use 30. This editing makes it possible to train a model on data of different lengths.

However, if the same kind of motion primitive is performed with a different velocity, it can change the model's ability to represent it with the same code. In the domain of signing data, the same signs can also be produced at different speeds, so that one motion primitive is produced within a different number of frames. This is supported by the difference we discovered in the types of codes a mixed model learns from different types of data, indicating that the current architecture needs different codes for different velocities. As a result, there is a limit to the current model's ability to learn and generalize well over different types of data, even when dealing with the exact same signs. This highlights the need to investigate the possibility to create an unsupervised setup that can capture time-invariant motion primitives for this task.

6.2. Generating New Samples

Given the limited amount of annotated sign language data, training an unsupervised model that can be used for a downstream task such as sign language production is of great interest. Even though it is possible to directly sample from the codebook of our model, it yields human-like but nonsensical results. Training a class, or language, guided model for code generation could yield more interesting results but is left as future research.

Additionally, by observing sampled and reconstructed sequence data we identify some limitations of the setup such as a need to improve the finger tracking and also increase the expressiveness of the model. For examples of generated sequences we refer to our project page[3].

7. Acknowledgements

This work was supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation, by the Swedish Research Council (VR) proj. 2023-04548.

8. Bibliographical References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Carl Börstell. 2023. Extracting sign language articulation from videos with mediapipe. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Hezhen Hu, Weichao Zhao, Wengang Zhou, and Houqiang Li. 2023. Signbert+: Hand-model-aware self-supervised pre-training for sign language understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–20.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. 2023. Motiongpt: Human motion as a foreign language.

Johanna Mesch and Lars Wallin. 2015. Gloss annotations in the swedish sign language corpus. *International Journal of Corpus Linguistics*, 20(1):102–120.

Johanna Mesch, Lars Wallin, and Thomas Björkstrand. 2012. Sign Language Resources in Sweden: Dictionary and Corpus. In *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon [Language Resources and Evaluation Conference (LREC)]*, pages 127–130, Paris. European Language Resources Association (ELRA).

Ali Razavi, Aaron van den Oord, and Oriol Vinyals. 2019. Generating diverse high-fidelity images with vq-vae-2.

Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. 2022. Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11050–11059.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2018. Neural discrete representation learning.

Pan Xie, Qipeng Zhang, Zexian Li, Hao Tang, Yao Du, and Xiaohui Hu. 2022. Vector quantized diffusion model with codeunet for text-to-sign pose sequences generation. *arXiv preprint arXiv:2208.09141*.

Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. 2023. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4210–4220.

Payam Jome Yazdian, Mo Chen, and Angelica Lim. 2022. Gesture2vec: Clustering gestures using representation learning methods for co-speech gesture generation. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3100–3107. IEEE.

Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. 2020. Mediapipe hands: On-device real-time hand tracking.

[3] www.speech.kth.se/research/vq-sign

Zhenxing Zhou, Vincent WL Tam, and Edmund Y Lam. 2021. Signbert: a bert-based deep learning framework for continuous sign language recognition. *IEEE Access*, 9:161669–161682.

9. Language Resource References

Svenskt teckenspråkslexikon. 2024. *Swedish Sign Language Dictionary online*. Sign Language Section, Department of Linguistics, Stockholm University.