From Scarcity to Understanding: Transfer Learning for the Extremely Low Resource Irish Sign Language

Ruth Holmes
SFI Lero – Trinity College Dublin
holmesru@tcd.ie

Ellen Rushe
SFI Lero – Trinity College Dublin
ellen.rushe@tcd.ie

Mathieu De Coster
IDLab-AIRO – Ghent University – imec
mathieu.decoster@ugent.be

Maxim Bonnaerens
IDLab-AIRO – Ghent University – imec
maxim.bonnaerens@ugent.be

Shin'ichi Satoh
National Institute of Informatics
satoh@nii.ac.jp

Akihiro Sugimoto
National Institute of Informatics
sugimoto@nii.ac.jp

Anthony Ventresque
SFI Lero – Trinity College Dublin
anthony.ventresque@tcd.ie

Abstract

One of the most significant challenges to sign language recognition (SLR) today is the low resource nature of sign language datasets, with many datasets being extremely low resource. Transfer learning is therefore a promising, and likely indispensable, method of increasing recognition performance. The use of pose estimation models, which are typically trained on a large and diverse population, can also aid generalization for extremely low resource sign languages. However, research on transfer learning for pose estimation keypoints as inputs has been limited. In this work, we explore transfer learning as a means to improve SLR classification performance for the extremely low resource Irish Sign Language (ISL). We show that transfer learning on larger datasets containing secondary sign languages significantly improves performance on our target sign language, ISL. To understand these results and the attributes that make one dataset better than another for pre-training, we analyse the linguistic relationships between these datasets. We find that certain attributes of datasets are associated with better transfer learning performance. We hope that our findings will not only motivate further research into transfer learning for pose keypoint-based SLR but also act as a practical guide to researchers on choosing the most suitable datasets with which to pre-train models.

1. Introduction

Signed Languages (SL) are the main form of communication for the Deaf and Hard-of-Hearing community. It is not a standalone international language but a collective term for many languages using a visual-gestural modality of communication. The World Federation of the Deaf reports over 200 SLs used by approximately 70 million people worldwide [1]. Within this exists many regional and dialectical differences and, like spoken languages, the manner in which these languages have travelled and developed over time has also resulted in many traceable historical influences. For example, the "French Sign Language Family" encompasses many modern European SLs such as Dutch, Italian and Irish SLs as well as American Sign Language (ASL) [20, 24]. These linguistic relationships introduce an interesting lens through which to view transfer learning strategies for Sign Language Recognition (SLR), particularly in the context of extremely low resource languages such as Irish Sign Language (ISL). More specifically, the areas where the attributes of different SLs overlap, such as hand-shapes, movements, or even entire signs, could provide an important bridge for knowledge transfer.

In this paper we will focus on one extremely low resource language in particular, ISL. ISL is the primary mode of communication for approximately 5,500 individuals and is estimated to have a total of just 60,000 users (hearing and

[1] https://wfdeaf.org/our-work/

Deaf) [20]. At the time of writing, there are just two ISL datasets curated for research purposes: The Signs of Ireland (SoI) [19] and the Irish Sign Language Hand-Shape (ISL-HS) dataset [28][2]. Notably, the latter is a very small dataset comprising of images of the 26 ISL alphabet handshapes. However, it is the only ISL dataset publicly available for download [3]. SoI includes continuous sign language data collected from Deaf ISL users across Ireland by fellow Deaf community member Deirdre Byrne-Dunne. As a result, the data is uniquely "natural" in the lexical choice and manner of the participants [19].

Despite the high quality of this dataset, it is still relatively small, with a limited vocabulary, making it challenging to use for SLR. To the best of our knowledge, there are no other signer independent [4] state-of-the-art SLR results reported for this dataset. Consequently the first aim of this paper is to address this gap in the literature by providing a baseline for SLR performance on this dataset. Furthermore, given the limited data that is available, we hypothesise that performance could be significantly improved by transfer learning. Therefore, our second aim is to explore transfer learning from other, larger SL datasets as a means to improve SLR for ISL. We explored a number of different datasets and found that some of the datasets we use for knowledge transfer were more helpful than others. Conventional wisdom would suggest that the largest dataset will necessarily provide the largest boost in performance. Surprisingly, we find that this is *not* the case, with more specific linguistic features being more influential on performance. We therefore explore the effect of the etymological closeness of SL datasets on their effectiveness for transfer learning. We hope that this analysis can act as a useful tool for practitioners when attempting to identify the best dataset to use for pre-training when fine-tuning on a extremely low-resource dataset.

The remainder of the paper[5] is structured as follows: Current work in the field of SLR is described in Section 2, specifically those using pose estimation based pipelines; We describe our method of performing transfer learning and our approach to the analysis of linguistic features in Section 3. Section 4 describes the data, model and implementation details; In Section 5 we present the results of these experiments; Finally, Section 6 concludes with a discussion of our findings and suggestions for potential future work.

[2]There is also an ongoing initiative by Dublin City University to produce a glossary of STEM related terms for ISL which could not be included at the time of this submission https://www.dcu.ie/islstem.

[3]https://github.com/marlondcu/ISL

[4]A data configuration whereby there are different individuals in each of the training, validation and test sets.

[5]This work was supported, in part, by SignON, a project funded by the European Union's Horizon 2020 Research and Innovation programme under grant No. 101017255; and by Science Foundation Ireland grant 13/RC/2094 P2 to Lero.

2. Related work

The lack of large-scale, diverse datasets is one of the major challenges associated with SLR research [8, 1, 10, 15]. Coupled with the data-hungry nature of deep learning-based machine learning techniques which comprise much of the state-of-art, this scarcity presents a critical obstacle to the development of translation systems capable of functioning in real-world signing scenarios. In fact, most SLs are extremely low resource, with multiple studies experimenting on datasets with as few as six individual signers [27, 9, 28, 29]. This can even be the case for popular SLR datasets, such as the RWTH-PHOENIX-Weather corpus which includes just seven individuals and a vocabulary of just 911 signs [11] relative to the several thousands that compose a SL. For models trained on these kinds of very limited datasets, the risk of bias propagation is significant, particularly in cases where raw images are used as input. Holmes *et al.* [15] found that this can be somewhat mitigated by the use of pose estimation keypoints rather than raw images. Pose estimation keypoints are a popular input representation in many SLR works [26, 7, 16]. Aside from an immense reduction in dimensionality when compared to image-based representations, the scale and variety of the training data used in the development of these pose estimation models lead to a greater level of invariance to to different visual conditions [26]. Furthermore, pose estimation frameworks such as Google's MediaPipe [22] publish accompanying details of the individuals it has been trained on [25], which suggests that these models have a greater level of robustness to human variation than could ever be achieved using a small SLR dataset with a very limited number of participants. Given these advantages, and to ensure the best possible generalisation beyond the evaluation sets of our the datasets we use, this paper will use models trained on pose estimation keypoints for pre-training.

Some recent works have suggested that pre-training on secondary more well-resourced SL data can improve performance on smaller target sign language datasets. Sharma *et al.* [30]. for instance, have made use of transfer learning from isolated to continuous Indian Sign Language data. Notably, however, the authors used sensor data as input which is a significantly more invasive data collection approach for participants than models using pose estimation keypoints or images as input. Similarly Bird *et al.* [3] use weights learned from British Sign Language (BSL) gestures to improve classification performance on a smaller ASL dataset. Here, the authors used a late fusion of image and "bone" data (captured using Leap Motion) classification models. However, works pertaining to knowledge transfer between SL datasets using pose estimation based architectures [33, 6] are limited. Instead, it is common practice [26, 7, 21] for these models to be trained from scratch resulting in an early plateau in performance [6].

3. Transfer Learning for Pose-based SLR

Transfer learning has become a cornerstore of modern computer vision models, so much so that pre-trained image models are a standard offering in the most popular deep learning frameworks such as PyTorch [6] and Tensorflow [7]. Given the low-resource nature of SLR for ISL, transfer learning will mostly likely be an essential component of a well-performing SLR model for ISL. However, pretraining on pose estimation keypoints has not been explored as extensively as transfer learning between models trained on images. This means that the types of features that are typically transferred from one pose-based task to another, and therefore the most useful characteristics of the pre-training keypoint datasets, remains largely unexplored. This gap in the literature makes it difficult to ascertain the most effective data to use for pre-training.

One approach to pre-training would be to train on vast amounts of data, à la large scale computer vision and language models, with the general consensus being that, the larger the dataset, the more performance that can be gained. This may well be true when the smaller target dataset contains some components of the larger dataset such as similar demographics of people, geographical location or overlapping set of classes. The idea of a truly generalised dataset, however, is largely a myth as no dataset, however large, can be free of distributional bias of some description [5]. In reality, the question that is typically more useful is "What dataset will provide us with the most attributes that overlap with our target dataset?". This is the question which we seek to answer in our experiments, in particular with respect to ISL. We do so by performing the following analysis:

Pre-training Dataset: To determine the dataset that is most effective for transfer learning, we perform pre-training on each dataset and evaluate the performance of the resultant model when fine-tuned on the low resource target dataset.

Degree of Fine-tuning: In this step we seek to determine the effect of fine-tuning all layers of the network versus fine-tuning the final classifier layer alone.

Gloss Analysis: Though pre-training models and evaluating their effect on performance on the target dataset is the most obvious approach to determine the most effective dataset, it would also be useful if we could establish the most appropriate dataset without the need to pre-train several models. We therefore perform analysis on the glosses [8] and discover the degree to which the vocabulary and lexical structure within them overlaps between the datasets used for pre-training and the ISL dataset used. The purpose of this analysis is to establish whether we can choose the most appropriate pre-training datasets a priori without having to train a candidate model. This step also provides us with an insight into the attributes of each dataset that make it useful for pre-training in the target SL.

[6] https://pytorch.org/

[7] https://www.tensorflow.org/

[8] Glosses are the labels assigned to each sign in written language. This is the primary means of annotation for sign language datasets.

4. Experimental Setup

This section will first detail the datasets used in our experiments along with the preprocessing performed. Next we will detail the model used for pre-training and fine-tuning. Finally we will provide implementation details for our gloss analysis.

4.1. Data

We experiment with a number of larger-resource sign language datasets in order to determine the most effective dataset with which to pre-train. Specifically, we attempt to transfer from two ASL datasets and one Flemish Sign Language (Vlaamse Gebarentaal, VGT) dataset. In this section, we briefly describe each dataset.

4.1.1 ISL

The Signs of Ireland dataset [19] is used to create the ISL dataset for these experiments. Specifically, the "Personal Story" and "Frog Story" activities are used. This dataset exhibits a long-tailed class imbalance with the majority class making up 9.5% of the total number of samples. Additionally, 24.1% of these samples are comprised of pointing/directional signs and basic gestures (i.e. those gloss labels not pertaining to distinct lexical items).

4.1.2 VGT

Corpus VGT [32] is a continuous Flemish SL dataset curated for linguistic research and is used here to construct the VGT dataset for these experiments. The resulting dataset also exhibits a long-tailed distribution with the majority class (a pointing sign) making up 10.4% of the total samples.

4.1.3 MS-ASL

The MS-ASL [17] dataset is an isolated sign language dataset and is publicly available here [23]. It is a collection of educational videos scraped from the internet however, at the time of downloading, many of these had been removed/made unavailable. Those that remained were downloaded and processed in the same manner as the above. Though the class distribution is not entirely balanced, it is

not to the same extent as the other datasets in this work. After preprocessing, the majority class constitutes just 0.5% of the total number of samples, with the fewest being 0.16%.

4.1.4 Google ASL

In our experiments we also use data taken from Google's recent Isolated Sign Language Recognition Kaggle competition [12]. Similar to MS-ASL, the class imbalance is not as extreme with the majority class making up 0.4% of the total samples. In terms of SLR, this dataset is extremely large. We collected a total of 94,477 processed samples, making it almost four times larger than the next largest dataset included here. This dataset is unique in that it contains recordings of one-handed signing, because participants needed to hold their smartphone (the recording device) with their other hand. As all of the datasets that we use in this paper feature two-handed signing, this may influence the transfer learning performance.

Table 1. Dataset statistics after preprocessing.

Dataset	# Samples	# Classes	# Participants
ISL	4,013	224	37
VGT	24,967	292	111
MS-ASL	12,259	402	173
Google ASL	94,477	250	21

In the case of ISL, VGT and ASL datasets, samples are curated based on the available gloss annotations. Larger videos are broken down into a collection of these word-level clips ranging from 0.012 to 10.12 seconds in length. Glosses relating to finger-spelled items are excluded. In the case of Google's Kaggle competition data, samples were provided in the form of pre-extracted MediaPipe keypoints relating to sequences ranging from 2 to 537 frames in length.

A stratified (on label) and grouped (on signer) split was performed to ensure that: 1. the class distribution is similar in the training, validation and test sets; 2. the data configuration is signer independent. We also ensure that each sign occurs at least once in the training and validation subsets. The number of included samples, classes and participants in each of these datasets is summarised in Table 1.

4.2. Data Preprocessing

Clips of ISL, VGT and MS-ASL datasets, were processed using MediaPipe Holistic [13] to extract 67 keypoints associated with the hands and upper body of each participant. In cases where hands could not be detected, temporal imputation (linear interpolation) is performed to naively infer these missing keypoints. Data normalisation is also performed where we shift to the centre of the chest and scale such that the distance between the shoulders is one. A sequence of these processed keypoints form the input to the model. The video data in Google's ASL dataset were already processed with MediaPipe Holistic: no video data were available but only keypoints were provided. We select the same 67 keypoints and process them in the same way as the other datasets.

4.3. Model

The model architecture used here is comprised of five stages. In the first of these, local temporal patterns are learned for each input feature (i.e., every coordinate individually) using residual depthwise 1D convolutions. We stack four 1D convolutional layers with increasing kernel size (3, 5, 7, 9) and add padding to maintain the sequence length. Secondly, non-linear relationships between individual keypoint coordinates are learned from embeddings generated independently from each frame in the sequence. The frame embedding consists of four blocks, each of them containing a linear layer followed by layer normalisation [2] and the GELU activation function [14] and regularised with dropout [31]. The final block does not contain the GELU activation, but a residual connection that adds the output of the first stage (projected onto the same space). Next, local temporal patterns within this embedding sequence are detected with a limited receptive field using another stack of residual depthwise 1D convolutions similar to the first stage, except that every convolution block now consists of two convolutional layers with GELU activations in between. Following this, global temporal information is learned using self-attention in which the receptive field covers the entire sequence. Finally, the resulting vector is used as input to the final classification layer. The architecture is illustrated in Figure 1.

4.4. Training

Table 2 summarises the hyperparameters common to all versions. In each case, the model was trained for a maximum of 50 epochs with fine-tuning. However, early stopping is employed to monitor validation loss with a patience of 20. The Adam optimizer [18] is used with an initial learning rate of 0.0003. The learning rate is reduced on a plateau by a factor of 0.1, monitoring validation accuracy with a patience of 5 epochs.

Table 2. Common hyperparameter values used in the experiments.

Hyperparameter	Value
Batch size	64
No. attention layers	4
No. attention heads	8
Feature size	134
Embedding size	192

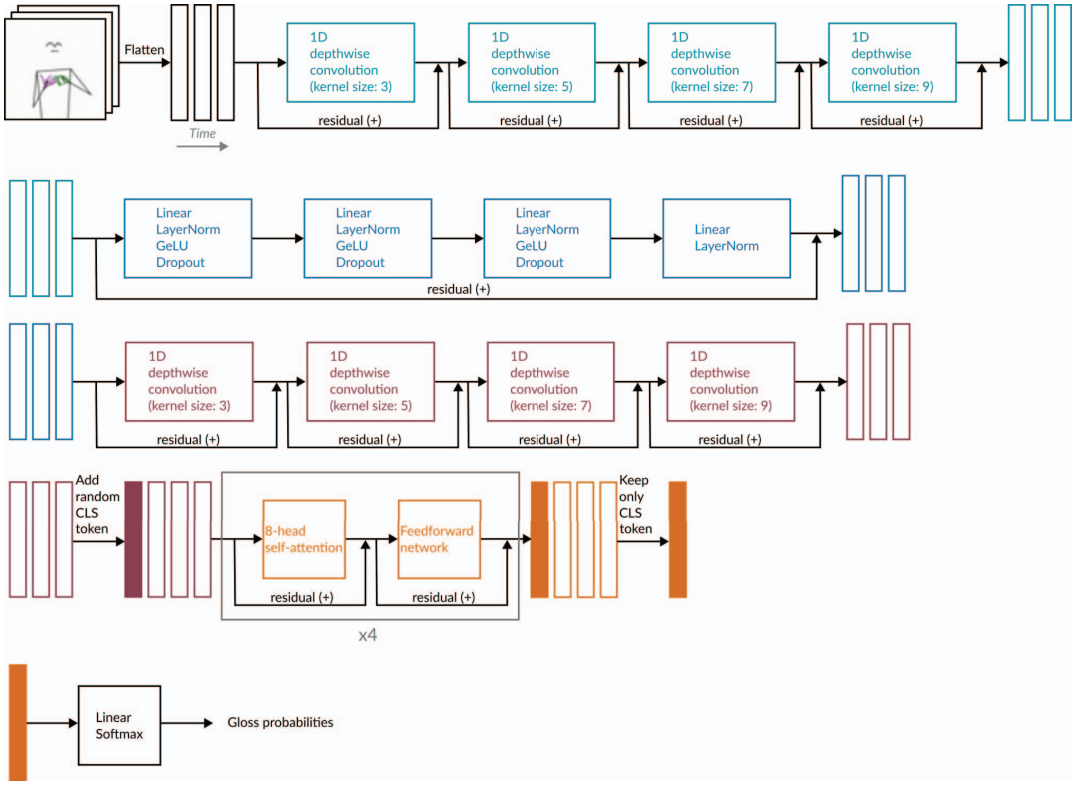

Figure 1. Diagram of the SLR model architecture.

4.5. Gloss Analysis

In this step we seek to gain an understanding of the similarities between datasets by analysing their glosses. Though glosses by no means provide a full description of the sign language being used, they can provide some insight into the lexical content of the datasets in the absence of a complete translation. We first perform some basic transformations to standardise the format of glosses to make them comparable. This process is somewhat complicated by the fact that different annotation conventions are used to convey gestures and different variations of signs in each dataset. For Corpus VGT, we remove any strings used to denote the particular "version" of a given gloss by removing the last portion of a gloss string separated by a '-' character, e.g. 'HAVE-A' to 'HAVE'. We then translated the written Flemish glosses to written English using Google Translate [9] and converted the glosses from all datasets to lowercase. The following comparisons are then performed:

Distribution of glosses: We first compare the distribution of overlapping glosses to determine how similar the frequency for these glosses are.

Distribution of lemmas: Next, we lemmatise the glosses to compare the distribution of lemmas. This step removes the effects of inflected forms of the written words used as glosses. Lemmatisation is performed using the NLTK [4] WordNet Lemmatizer [10].

Distribution of Part-Of-Speech tags: Finally, we obtain the Part-of-Speech (PoS) tags for each standardised gloss in order to compare the grammatical composition of glosses. PoS tagging is implemented using NLTK [11].

Overlapping terms vs. all terms For each of the aforementioned comparisons, we compare the distributions of glosses, lemmas and PoS tags that overlap between SoI and each other dataset. In the final step, we perform this same analysis on all glosses combined for each dataset paired with SoI.

It is important to note here that the purpose of this analysis is to give us a general indication of the "closest" dataset to SoI and therefore this analysis is by no means exhaustive. There are a number of limitations to this analysis as a result. For instance, in some cases glosses were written as a compound words, combining different terms in a way that would

[9] https://translate.google.com/

[10] https://www.nltk.org/_modules/nltk/stem/wordnet.html

[11] https://www.nltk.org/api/nltk.tag.pos_tag.html

not typically be done in written English. Furthermore, since different annotations conventions are used in each dataset, gestures that cannot be directly mapped to an equivalent form in written language are annotated in different ways. In this work, we have not grouped different gesture annotations and leave more complex analysis of gestures to future work. Nevertheless, despite these limitations, we found that the basic analysis outlined above mapped well to the results achieved by models in the transfer learning experiments.

5. Results

A summary of the baseline F1-scores for each of the datasets included in these experiments can be found in Table 3. These scores are calculated using the model state from the best performing epoch in terms of validation accuracy. The weights from these Google ASL, MS-ASL and VGT models are then used as the basis for transfer learning to ISL.

Table 3. Baseline F1-scores for each dataset.

Dataset	Training	Validation	Test
ISL	0.7604	0.243	0.2221
Google ASL	0.9052	0.6591	N/A [a]
MS-ASL	0.9983	0.6545	0.6305
VGT	0.7249	0.5007	0.482

[a] There are no results for the Google ASL test subset as data was not released at the time of experiments.

We will first look at the effect of the pre-training dataset used on classification performance. Next, we will evaluate the extent to which fine-tuning increases performance. Finally, we will detail the results of our analysis on the dataset gloss annotations and provide a discussion on how the outcomes can be leveraged by researchers in future.

5.1. Pre-training Dataset

A summary of the results of each transfer learning strategy can be found in Table 4. In all cases, the performance of ISL benefits significantly from pre-training.

Despite having far more samples than the other datasets, pre-training on Google's ASL dataset does not yield the best results. We see two possible reasons for this: 1. there is a domain mismatch between the isolated signing in this dataset and the coarticulated signing in the SoI dataset; 2. there is another domain mismatch between the one-handed signing in this dataset and the two-handed signing in the SoI dataset. In fact, MS-ASL has a fraction of the samples of Google's ASL dataset, yet the SoI performance after pre-training on MS-ASL is similar. By far the best performance is achieved when pre-training on the Corpus VGT. This is most likely because Corpus VGT also features coarticulated signing. This insight is notable as it suggests that pre-training on coarticulated signing is crucial to properly leverage transfer learning for real-world signing.

Table 4. Summary of results (F1 scores) for ISL for each transfer learning strategy *with* fine-tuning.

Transfer	Training	Validation	Test
Google ASL	0.8789	0.2809	0.2508
MS-ASL	0.9638	0.2776	0.2483
VGT	0.9073	**0.3311**	**0.2736**

5.2. Effect of Fine-tuning

To evaluate the extent to which fine-tuning increases performance, we also created a version of these experiments where all but the final classification layer of the model is frozen. A summary of these results can be seen in Table 5. Interestingly, the significantly negative impact of the Google ASL model in this case further highlights the presumed domain mismatch. Meanwhile, the Corpus VGT continues to be of significant benefit.

Table 5. Summary of results (F1 scores) for ISL for each transfer learning strategy *without* fine-tuning.

Transfer	Training	Validation	Test
Google ASL	0.4312	0.1527	0.1807
MS-ASL	0.5259	0.2085	0.228
VGT	0.6392	**0.2687**	**0.2483**

5.3. Gloss analysis

Here, we attempt to understand the attributes of a given dataset that makes it more beneficial for pre-training than others. We hope that this can not only give us an indication of the most salient types of features for the SLR models but also provide practitioners with concrete measurements that can we used to guide their choice of pre-training dataset.

First, we look to the number of overlapping glosses and lemmas between SoI and the other datasets in order to establish whether these are, in any way, associated with the performance of models. In Table 6, we can see that the number of overlapping terms does not seem to be the primary factor influencing the difference in performance between pre-training datasets. In fact, MS-ASL, which leads to the worst performance of the three datasets in terms of F1-score, has the most overlapping glosses and lemmas. It is therefore useful to look more closely at the distribution of these overlapping terms to determine whether this reveals the same pattern of performance increase/decrease between datasets.

Figures 2 and 3 show the distribution of overlapping terms between SoI and MS-ASL/Corpus VGT, the worst

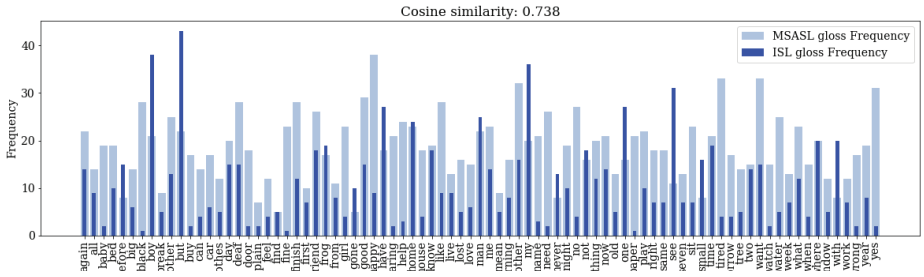

Figure 2. Distribution of glosses that overlap between SoI (ISL) and MS-ASL

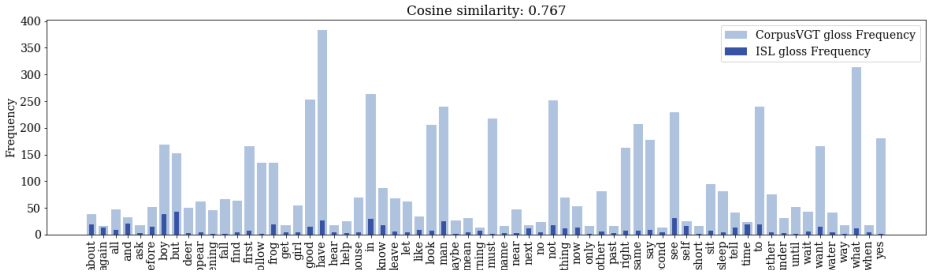

Figure 3. Distribution of glosses that overlap between SoI (ISL) and Corpus VGT.

Table 6. Number of overlapping glosses/lemmas/PoS tags between datasets.

	Overlapping glosses	Overlapping lemmas	Overlapping PoS tags
ISL-MSASL	83	83	16
ISL-GoogleASL	55	55	11
ISL-CorpusVGT	66	68	16

Table 7. Cosine similarity between distributions of overlappling gloss/lemma/PoS tag frequency for ISL and each other dataset.

	Gloss	Lemma	PoS tag
ISL-MSASL	0.738	0.739	0.987
ISL-GoogleASL	0.754	0.754	0.994
ISL-CorpusVGT	0.767	0.766	0.995

and best performing pre-training datasets respectively. Table 7 provides the cosine similarity between the gloss, lemma and PoS tag frequency for overlapping terms in order to quantify the similarity between these distributions. We can see that the distribution between overlapping terms in Corpus VGT is closest to SoI for glosses, lemmas and PoS tags. In fact, the ranking of the models in terms of F1-score matches the ranks of the similarity scores exactly for each of these datasets. This suggests that the frequency distribution of the overlapping glosses, lemmas and PoS tags may be indicative of the benefit a given dataset has for pre-training before fine-tuning on a particular target dataset.

Analysing the overlapping glosses, however, ignores the presence of *non-overlapping* glosses which will inevitably affect the training of pre-trained models and, consequently, affect the overall suitability of a dataset for use in pre-training. In Table 8, we show the cosine similarity between SoI and the other three datasets in terms of the frequency distribution of the union of their glosses, lemmas and PoS tags. Clearly, here we can see that the number of non-overlapping glosses has a large effect on the overall similarity between frequency distributions of glosses and, therefore, lemmas. This is due to the fact that there are a large number of glosses that are not in common between pairs of datasets. As a matter of fact we see that for glosses and lemmas, when ranking the datasets in terms of this type similarity to SoI, the ranks are inversely proportional to the F1-score.

Table 8. Cosine similarity between distributions of gloss/lemma/PoS tag frequency for ISL and each other dataset.

	Gloss	Lemma	PoS tag
ISL-MSASL	0.186	0.187	0.987
ISL-GoogleASL	0.126	0.126	0.990
ISL-CorpusVGT	0.075	0.076	0.994

However, we have yet to discuss the role of PoS tags, which we see consistently rank the datasets in terms of cosine similarity proportionally to their ranks in terms of F1-score. Figures 4 and 5 show the distributions of PoS tags for Corpus VGT and MS-ASL, the best and worst performing datasets respectively for pre-training. This indicates that having a similar frequency of certain grammatical structures is associated with better pre-training performance. When

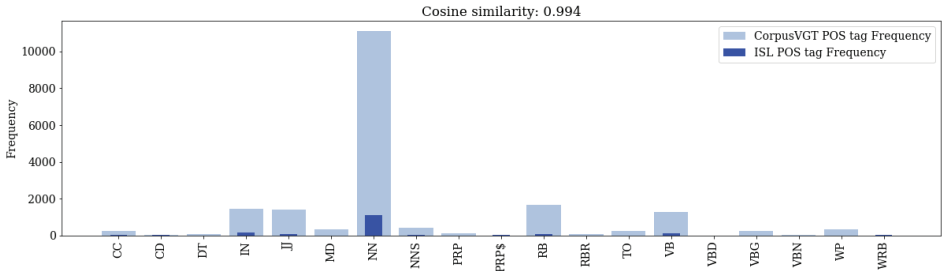

Figure 4. Distribution of PoS tags for all glosses in SoI (ISL) and Corpus VGT.

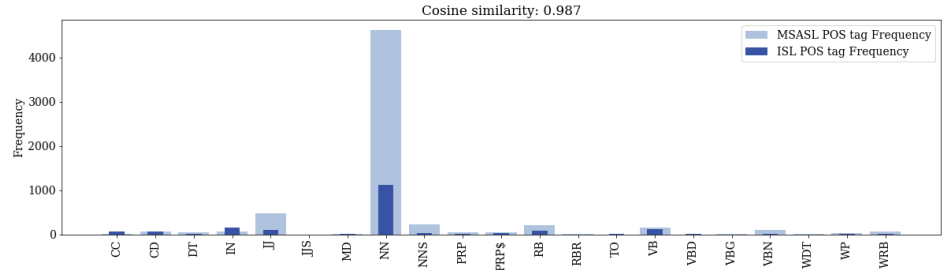

Figure 5. Distribution of PoS tags for all glosses in SoI (ISL) and MS-ASL.

we consider that Corpus VGT is composed of continuous signing, like SoI, (rather than isolated signs which compose MS-ASL and GoogleASL) the reasons for this variation in grammatical structure becomes clearer. The presence of some signs will clearly be more or less common in natural, continuous sign language than in a collection of isolated signs. The degree of difference between the frequency distributions, however, is not extremely large suggesting that it is a combination of factors that lead to increased performance. For instance, this could be due to a similar distribution of overlapping glosses, as discussed above, or some other factors that are more difficult to analyse due to their complexity. One such factor may be that there are more variations of particular signs in continuous sign language due to coarticulations or the presence of sign-language specific grammatical structures that are not discernible using a standard PoS tagger, given that these tools are developed for written languages, not sign languages.

In summary, our analysis shows that the similarity between the frequency distribution of overlapping glosses/lemmas and certain grammatical characteristics are associated with the performance of a given dataset as a pre-training dataset for SoI. We hope that this can guide practitioners when seeking to determine the suitability of datasets for use in pre-training and motivate the use of continuous sign language datasets for pre-training.

6. Conclusions

The lack of large, varied datasets is a key challenge in SLR, in particular for extremely low resources languages such as ISL. Transfer learning has therefore become an essential component of SLR systems. Though some work has been done in the area of transfer learning for raw-image-based SLR, there has been a distinct lack of evaluation of the effectiveness of transfer learning for models that use pose estimation derived from sign language videos. Pose estimation models are indispensable to these pipelines due to the fact that these models are trained on a more diverse population than would be available in any current SL dataset. It is therefore crucial to evaluate the efficacy of transfer learning using models that use pose estimation keypoints as input. In this work we have, to the best of our knowledge, provided the first signer-independent SLR results for the SoI dataset. Secondly, we have shown that pre-training on secondary sign language datasets provides a significant boost in recognition performance for keypoint-based model, in particular when end-to-end fine-tuning is done on the target sign language. Finally we have compared the characteristics of the gloss labels of these datasets to determine whether the suitability of datasets for pre-training can be established *a priori*. In doing so, we have revealed an association between the suitability of a dataset for pre-training and the similarity between their gloss and grammatical feature distributions to that of the target dataset. In future we aim to expand our evaluation to an extended set of datasets, additional architectures, and broaden our analysis to glosses and grammatical features that are more specific to SLs.

References

[1] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 35–53. Springer, 2020. 2

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 4

[3] Jordan J Bird, Anikó Ekárt, and Diego R Faria. British sign language recognition via late fusion of computer vision and leap motion with transfer learning to american sign language. *Sensors*, 20(18):5151, 2020. 2

[4] Steven Bird and Edward Loper. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain, July 2004. Association for Computational Linguistics. 5

[5] Abeba Birhane, Vinay Prabhu, Sang Han, and Vishnu Naresh Boddeti. On hate scaling laws for data-swamps. *arXiv preprint arXiv:2306.13141*, 2023. 3

[6] Mathieu De Coster, Ellen Rushe, Ruth Holmes, Anthony Ventresque, and Joni Dambre. Towards the extraction of robust sign embeddings for low resource sign language recognition. *arXiv preprint arXiv:2306.17558*, 2023. 2

[7] Mathieu De Coster, Mieke Van Herreweghe, and Joni Dambre. Isolated sign recognition from rgb video using pose flow and self-attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3441–3450, 2021. 2

[8] Mirella De Sisto, Vincent Vandeghinste, Santiago Egea Gómez, Mathieu De Coster, Dimitar Shterionov, and Horacio Saggion. Challenges with sign language datasets for sign language recognition and translation. In *Calzolari N, Béchet F, Blache P, Choukri K, Cieri C, Declerck T, Goggi S, Isahara H, Maegaard B, Mariani J, Mazo H, Odijk J, Piperidis S, editors. LREC 2022, 13th International Conference on Language Resources and Evaluation; 2022 June 20-25; Marseille, France. Paris: European Language Resources; 2022. 10 p.* European Language Resources Association, 2022. 2

[9] Marco Fagiani, Emanuele Principi, Stefano Squartini, and Francesco Piazza. Signer independent isolated italian sign recognition based on hidden markov models. *Pattern Analysis and Applications*, 18(2):385–402, 2015. 2

[10] Jérôme Fink, Benoît Frénay, Laurence Meurant, and Anthony Cleve. Lsfb-cont and lsfb-isol: Two new datasets for vision-based sign language recognition. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021. 2

[11] 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*, volume 9, pages 3785–3789, 2012. 2

[12] Google. Google - Isolated Sign Language Recognition. https://www.kaggle.com/competitions/asl-signs/overview. Accessed: 2023-07-10. 4

[13] Google. MediaPipe Holistic. https://google.github.io/mediapipe/solutions/holistic.html. Accessed: 2023-07-05. 4

[14] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 4

[15] Ruth Holmes, Ellen Rushe, Frank Fowley, and Anthony Ventresque. Improving signer independent sign language recognition for low resource languages. In *Proceedings of the 7th International Workshop on Sign Language Translation and Avatar Technology: The Junction of the Visual and the Textual: Challenges and Perspectives*, pages 45–52, Marseille, France, June 2022. European Language Resources Association. 2

[16] Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. Skeleton aware multi-modal sign language recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3413–3423, 2021. 2

[17] 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. 3

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[19] Lorraine Leeson. Moving heads and moving hands: Developing a digital corpus of irish sign language. *Ireland¡/i¿,¡ i*, pages 25–26, 2006. 2, 3

[20] Lorraine Leeson, John I Saeed, and Carmel Grehan. 18 irish sign language (isl). *Sign Languages of the World: A Comparative Handbook*, page 449, 2015. 1, 2

[21] 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 *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1459–1469, 2020. 2

[22] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019. 2

[23] Microsoft. MS-ASL American Sign Language Dataset. https://www.microsoft.com/en-us/download/details.aspx?id=100121. Accessed: 2023-07-05. 3

[24] Agnès Millet, Nathalie Niederberger, and Marion Blondel. 10 french sign language. *Sign languages of the world: A comparative handbook*, pages 273–316, 2015. 1

[25] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019. 2

[26] Amit Moryossef, Ioannis Tsochantaridis, Joe Dinn, Necati Cihan Camgoz, Richard Bowden, Tao Jiang, Annette

Rios, Mathias Muller, and Sarah Ebling. Evaluating the immediate applicability of pose estimation for sign language recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3434–3440, 2021. 2

[27] Pisit Nakjai and Tatpong Katanyukul. Hand sign recognition for thai finger spelling: An application of convolution neural network. *Journal of Signal Processing Systems*, 91(2):131–146, 2019. 2

[28] Marlon Oliveira, Houssem Chatbri, Ylva Ferstl, Mohamed Farouk, Suzanne Little, Noel E O'Connor, and Alistair Sutherland. A dataset for irish sign language recognition. In *IMVIP*, 2017. 2

[29] Marlon Oliveira, Houssem Chatbri, Suzanne Little, Ylva Ferstl, Noel E O'Connor, and Alistair Sutherland. Irish sign language recognition using principal component analysis and convolutional neural networks. In *DICTA*, pages 1–8. IEEE, 2017. 2

[30] S Sharma, R Gupta, and A Kumar. Continuous sign language recognition using isolated signs data and deep transfer learning. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–12, 2021. 2

[31] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. 4

[32] Mieke Van Herreweghe, Myriam Vermeerbergen, Eline Demey, Hannes De Durpel, Hilde Nyffels, and Sam Verstraete. Het Corpus VGT. Een digitaal open access corpus van videos and annotaties van Vlaamse Gebarentaal, ontwikkeld aan de Universiteit Gent ism KU Leuven. www.corpusvgt.be, 2015. 3

[33] Manuel Vázquez-Enríquez, Jose L Alba-Castro, Laura Docío-Fernández, and Eduardo Rodríguez-Banga. Isolated sign language recognition with multi-scale spatial-temporal graph convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3462–3471, 2021. 2