Active Learning for Multilingual Fingerspelling Corpora

Shuai Wang   SHUAI.WANG@STUDENT.UVA.NL
Eric Nalisnick   E.T.NALISNICK@UVA.NL
Informatics Institute
University of Amsterdam
Amsterdam, Netherlands

Abstract

We apply active learning to help with data scarcity problems in sign languages. In particular, we perform a novel analysis of the effect of pre-training. Since many sign languages are linguistic descendants of French sign language, they share hand configurations, which pre-training can hopefully exploit. We test this hypothesis on American, Chinese, German, and Irish fingerspelling corpora. We do observe a benefit from pre-training, but this may be due to visual rather than linguistic similarities.

Keywords: Active Learning, Transfer Learning, Multilingual Sign Language

1. Introduction

While there has been recent advances in technologies for written languages (e.g. *BERT* (Devlin et al., 2019), GPT-3 (Brown et al., 2020)), *signed languages* have received little attention (Yin et al., 2021). In turn, the globe's 70 million deaf people and 360 million hard-of-hearing people largely do not benefit from modern language technologies. One reason for this gap is that *sign language processing* (SLP) is inherently more challenging since identifying signs is a complex computer vision task—compared to string matching for written languages. Yet, the scarcity of sign language corpora is the largest hurdle to advances in SLP. There are, of course, relatively fewer users of signed languages, leading to less available training data. Moreover, the relative scarcity of native signers makes procuring high-quality labels and annotations for supervised learning difficult and costly.

In this work, we make progress on alleviating data scarcity in SLP by applying *active learning* (AL) (Settles, 2012) to multilingual *fingerspelling* corpora. *Fingerspelling* is a sub-task in sign language communication: the signer spells-out a word associated with a concept that does not have an associated sign, such as a proper noun. For example, if a signer wants to reference "GPT-3," the signer would make a hand shape for each letter 'G,' 'P', and 'T'. See Figure 1 for example images. In addition to detecting sign-less concepts, recognizing fingerspelling is important because the hand configuration for some concepts are the same as the first letter of the associated word. For instance, in American sign language, the concept of 'ready' is signed with the hands in the 'r'-configuration.

While fingerspelling recognition is a relatively easier task than full SLP, it can still be challenging for low-resource languages. Consider Irish sign language: it has just 5000 active deaf users, and to make matters more complicated, has variations depending on the gender of the signer. Collecting large, diverse data sets for languages such as this one requires great
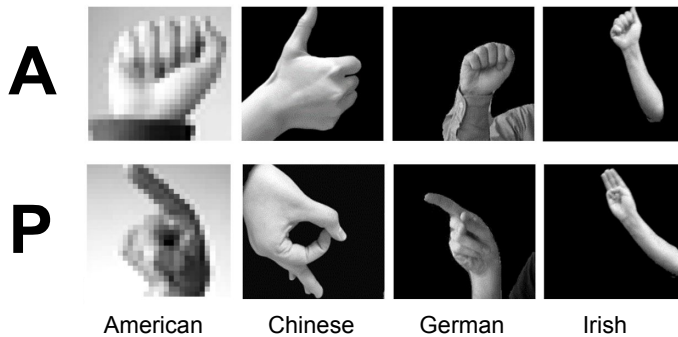
Figure 1: *Example Images.* The above figure shows the letters 'A' and 'P' from the ASL, CSL, GSL, and ISL data sets respectively.

effort. Thus, we also perform *transfer active learning* by exploiting linguistic relations across sign languages. Specifically, American, Irish, German, and Chinese sign languages all are descendants of French sign language, thus resulting in shared hand configurations, among other features. We conjecture that learning can be done more quickly by first pre-training a model on a related sign language and then performing AL on an especially low-resource target language.

We report two sets of experiments. In the first, we perform AL on American, Irish, German, and Chinese sign languages, demonstrating improvements over random sampling. In the second, we perform transfer AL by first pre-training the model on one of the other three sign languages. This is the first work to investigate leveraging linguistic relationships for transfer AL. While we observe some benefits, our initial results suggest that pairing data sets with a similar visual style improves transfer AL more than having a strong linguistic relationship but dissimilar visual style.

2. Data Sets: Multilingual Fingerspelling Corpora

We use the following publicly available fingerspelling corpora, all of which have a linguistic relationship due to being descendants from French sign language. In all cases, we standardize the image resolution to 28×28 pixels. For each data set, we resample the frequency of each letter so that it follows the character distribution of the corresponding written language. As a result, the data sets are class imbalanced.

1. **American Fingerspelling (ASL)**: This data set is availabe on Kaggle[1] and contains $34,627$ grayscale images (resolution 28×28) for 24 alphabetic characters used in American sign language (ASL). The characters 'J' and 'Z' are omitted because they are dynamic signs that cannot be well-represented by a static image. The images are closely cropped to the hand and in turn do not require pre-processing.

2. **Chinese Fingerspelling (CSL)** (Jiang et al., 2020): This data set consists of the alphabetic characters used in Chinese sign language (CSL). Specifically, there are 1320

1. https://www.kaggle.com/datasets/datamunge/sign-language-mnist

images, each labeled as one of 30 letters—26 are the characters A through Z and 4 are double syllable letters (ZH, CH, SH, and NG). The resolution is originally 256×256, and we down-sample them to 28×28. We discard the double syllable letters since they are unamenable to transfer learning. We convert all images to grayscale and remove 'J' and 'Z' so that the data set is aligned with the ASL data set.

3. **German Fingerspelling (GSL)** (Dreuw et al., 2006): This data set contains alphanumeric characters used in German sign language (GSL). The original data set consists of video sequences for 35 gestures: the 26 letters A to Z, the 4 German umlauts (SCH, Ä, Ö, Ü), and the numbers from 1 to 5. We extracted $20,904$ static images for the letters from the video frames. We remove the background environment using a ResNet-101 model pre-trained for segmentation (Chen et al., 2018). We again convert the images to grayscale and remove 'J' and 'Z'.

4. **Irish Fingerspelling (ISL)** (Oliveira et al., 2017): This data set contains the alphabetic characters used in Irish sign language (ISL). It consists of $58,114$ images for the 26 letters A through Z. This data set is unique from the others in that the sign is repeated by moving the forearm from a vertical to near horizontal position. Again, we convert the images to grayscale and remove 'J' and 'Z'.

Figure 1 shows images for the letters 'A' and 'P' from the four corpora listed above. The sign for 'A' is very similar for ASL, GSL, and ISL. The sign for 'P' is the same for ASL and GSL but different for CSL and ISL.

3. Methodology: Active Learning

Our aim is to obtain a predictive model for fingerspelling recognition. Given an input image, the model should be able to predict a label representing the alphabetic character being signed. Yet, given the resource constraints for signed languages (discussed in Section 1), we wish to train a highly accurate predictive model using as few labeled instances as possible. To achieve this, we rely on the methodology of *active learning* (AL) (Settles, 2012), which allows us to obtain a sequence of highly informative labels from an oracle (such as a human expert). The hope is that the intelligent selection of these labels allows the model to achieve satisfactory performance while minimizing queries of the oracle. Formally, we assume the model is first trained on an initial data set $\mathcal{D}_0 = \{\boldsymbol{x}_n, y_n\}_{n=1}^{N}$, with N being relatively small for the problem at hand. We also assume access to an unlabeled pool set $\mathcal{X}_p = \{\boldsymbol{x}_m\}_{m=1}^{M}$ such that $M >> N$. The oracle has access to the corresponding labels for this pool set, denoted $\mathcal{Y}_p = \{y_m\}_{m=1}^{M}$.

AL for image data has been well-studied (Gal et al., 2017), and various acquisition functions have been proposed. In preliminary experiments, we studied the following: maximum entropy (Shannon, 1948), *BALD* (Houlsby et al., 2011), variation ratios (Freeman, 1965), and mean standard deviation. We found all methods to be comparable but with variation ratios to have a slight edge. In turn, we use variation ratios in all experiments. At time step t, a point from the pool set is selecting according to:

$$\boldsymbol{x}_t^* \;=\; \underset{\boldsymbol{x} \in \mathcal{X}_{p,t-1}}{\arg\max} \quad 1 - \max_{\mathrm{y}} p\left(\mathrm{y}|\boldsymbol{x}, \boldsymbol{\theta}_{t-1}\right) \tag{1}$$

where $p\left(\mathrm{y}|\boldsymbol{x},\boldsymbol{\theta}_{t-1}\right)$ is the model's maximum class confidence after being fit during the previous time step. After obtaining $\boldsymbol{x}_t^*$, the oracle is queried for the associated label, and the model is re-fit to data that includes the selected feature-label pair (to obtain $\boldsymbol{\theta}_t$).

Transfer Active Learning Knowledge gained from other domains can be helpful to improve performance on the target domain. In fact, Tamkin et al. (2022) suggest that the ability to actively learn is an emergent property of pre-training. We combine this inspiration with the fact that ASL, CSL, GSL, and ISL are linguistically related, proposing to pre-train on a related fingerspelling corpora before performing AL on the target domain. Specifically, we assume access to an auxiliary data set $\mathcal{D}_{-1} = \{\boldsymbol{x}_l, y_l\}_{l=1}^L$. We train the model of interest on this data set before training it on the pool set $\mathcal{D}_0$. $\mathcal{D}_{-1}$ is seen by the model only once and not retained as part of the growing data set as AL proceeds.
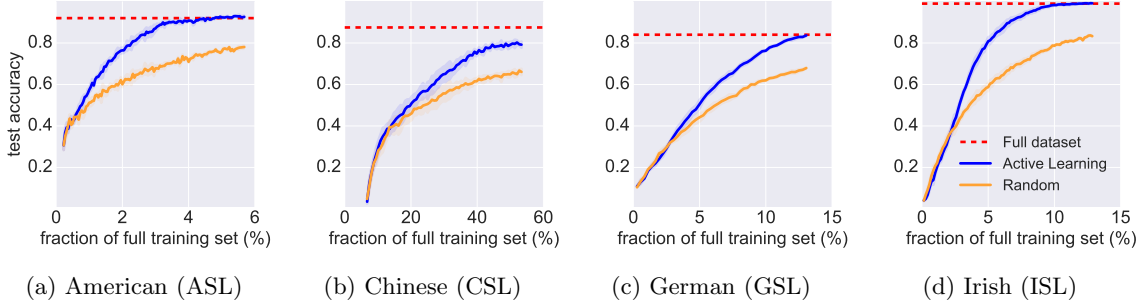

Figure 2: *Single Corpus Active Learning*: The figure above shows AL results for each fingerspelling corpus. Using variation ratios (blue) as an acquisition function is clearly superior to random sampling (yellow).

4. Experiments

We perform two types of experiment. For the first, we verify the effectiveness of AL for an individual fingerspelling corpus. For the second, we examine if AL can be made more sample efficient by pre-training on a linguistically related corpus. The following settings apply to all experiments. We construct the initial set $\mathcal{D}_0$ by choosing two samples per class uniformly at random ($N = 48$). For constructing test sets, we hold-out 30% of the data in CSL (because of its small size) and 10% for the other corpora. The remaining data is the pool set. The query size for each data set is different due their different sizes: 50 for ISL and GSL, 10 for ASL, and 5 for CSL. The classifier is retrained from a random initialization whenever new labels are procured. The classifier is trained for 50 epochs using a batch size of 128 with learning rate of 10^{-3}.

4.1 Single Corpus Active Learning

Implementation Details We compare against uniformly random acquisition as the baseline. We re-run experiments with three unique random seeds to account for sampling variation. The classifier is a neural network with two convolutional layers (ReLU activations
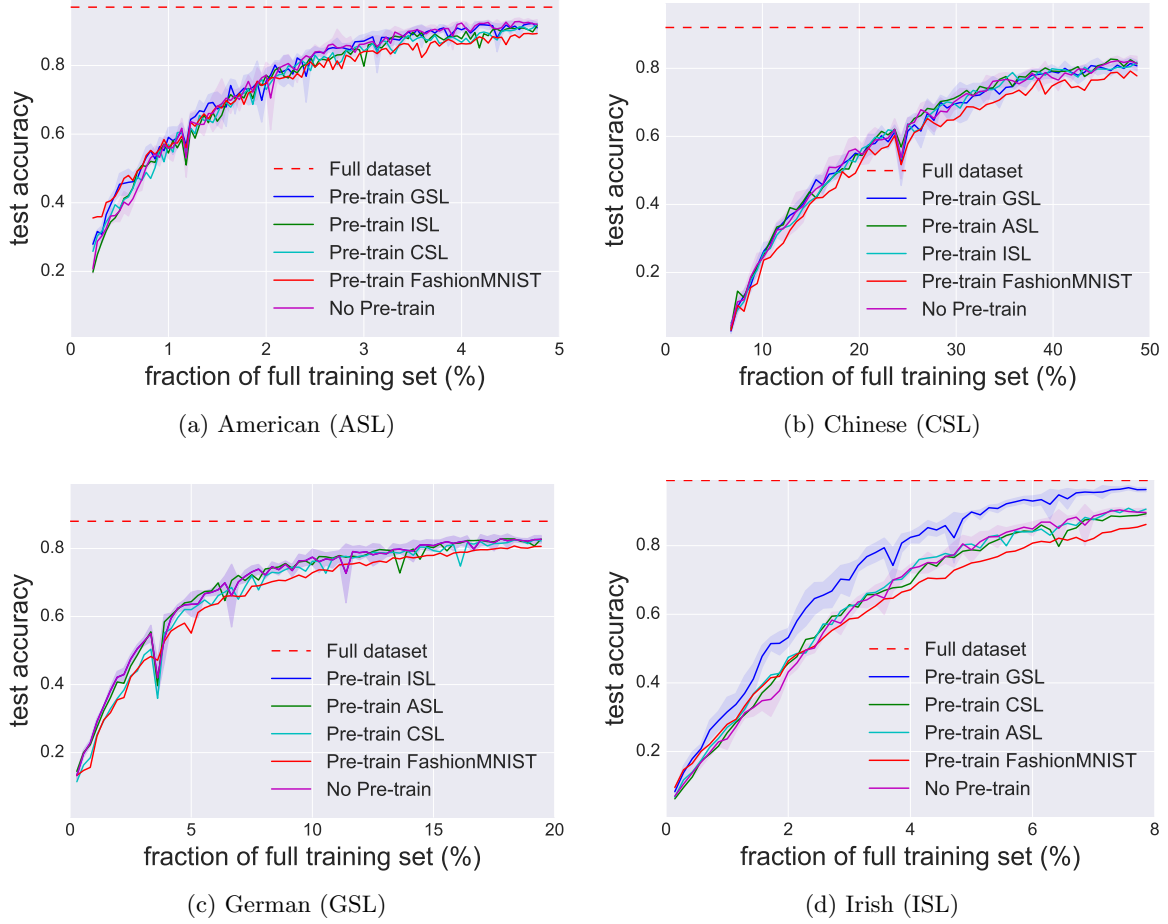
Figure 3: *Transfer Active Learning*: The figure above shows AL results for each finger-spelling corpus with pre-training and no pre-training. Pre-training with a non-fingerspelling corpus degrades performance in all cases.
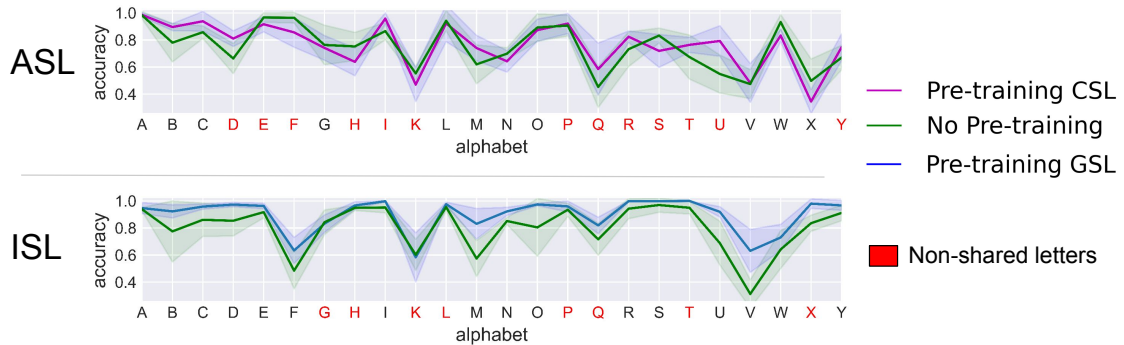

Figure 4: *Per Class Results for Transfer Active Learning*: Generally, we see the gains from pre-training are had by shared letters (red) but not exclusively ('X' in ISL is not shared by GSL).

and dropout with $p = 25\%$) and two fully-connected layers (ReLU activations and dropout with $p = 50\%$).

Results Figure 2 reports the results, showing test set accuracy (y-axis) as more data is collected (x-axis). We see a clear improvement over random sampling in all cases. Moreover, AL reaches near full-data-set performance by seeing just 15% (or less) of the original data in three out of four cases (ASL, GSL, and ISL). GSL sees the least benefit from AL, but surprisingly, we see that with just 12% of the training data, the model starts to surpass full-data-set performance. We suspect that this is due to outliers (created by frame extraction and/or cropping mentioned in section 2) that the AL method has not yet seen but detract from performance.

4.2 Transfer Active Learning

Implementation Details We next turn to the pre-training experiment. We use a ResNet-18 (He et al., 2016) as the backbone with the same structure as the previous experiment. We pre-train the classifier on the full data set from one corpus and then load the backbone of the pre-trained model with a re-initialized classification head. The entire structure of the neural network is retrained during AL. We re-run all experiments five times with different random seeds. As a control, we also perform a run with the model pre-trained on FashionMNIST (Xiao et al., 2017), another grayscale data set.

Results Figure 3 reports test accuracy for all pre-training configurations, including no pre-training. The only noticeable trend that generalizes to all corpora is that pre-training on FashionMNIST degrades performance in comparison to pre-training on a fingerspelling corpus *and* using no pre-training. While this observation supports our motivating hypothesis for transfer AL, otherwise, we see modest gains with pre-training. The only clear improvement over no pre-training is demonstrated for ISL pre-trained on GSL (subfigure d, blue line). As these are the only two data sets that include the signer's forearm, we suspect that visual similarly is causing the benefit rather than linguistic relationship, or else we would have seen improvements in other data pairings. Yet, curiously, this benefit is not symmetric since pre-training with ISL did not improve GSL performance more than pre-training with other corpora. Figure 4 shows the accuracy per class (letter). To make the performance gap more obvious, we show performance at the time step with the largest gap ($t = 40$). We expect to see the pink and blue lines (pre-training) to be higher than the green line (no pre-training) for letters shown in black (shared across languages). This is generally the case, but we also see improvements for non-shared letters, such as for 'X' in ISL.

5. Conclusion

We have reported the first transfer AL results for fingerspelling corpora. We observed a clear success (pretraining on GSL and performing AL on ISL), but this was not observed for other pairings nor is the ISL-GSL relationship symmetric. In future work, we plan to more carefully control for linguistic and visual relationships between data sets to better isolate the cause of the performance gains.

Acknowledgements

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 2020.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision*, 2018.

David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 1996.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

Thang B. Dinh, Van B. Dang, Duc A. Duong, Tuan T. Nguyen, and Duy-Dinh Le. Hand gesture classification using boosted cascade of classifiers. In *International Conference onResearch, Innovation and Vision for the Future*, 2006.

Philippe Dreuw, Thomas Deselaers, Daniel Keysers, and Hermann Ney. Modeling image variability in appearance-based gesture recognition. In *ECCV Workshop on Statistical Methods in Multi-Image and Video Processing*, 2006.

R. Elakkiya. Machine learning based sign language recognition: a review and its research frontier. *Journal of Ambient Intelligence and Humanized Computing*, 2021.

R. Elakkiya and K. Selvamani. An active learning framework for human hand sign gestures and handling movement epenthesis using enhanced level building approach. *Procedia Computer Science*, 2015.

Linton G. Freeman. Elementary applied statistics for students in behavioral science. In *Social Forces*, 1965.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep Bayesian active learning with image data. In *Proceedings of the International Conference on Machine Learning*, 2017.

Ruqian Hao, Khashayar Namdar, Lin Liu, and Farzad Khalvati. A transfer learning–based active learning framework for brain tumor classification. *Frontiers in Artificial Intelligence*, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Neil Houlsby, Ferenc Huszar, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv*, 2011.

Xianwei Jiang, Bo Hu, Suresh Chandra Satapathy, Shuihua Wang, and Yudong Zhang. Fingerspelling identification for Chinese sign language via AlexNet-based transfer learning and adam optimizer. *Scientific Programming*, 2020.

Ozge Mercanoglu and Hacer Keles. Autsl: A large scale multi-modal Turkish sign language dataset and baseline methods. *IEEE Access*, 2020.

Kathiravan Natarajan, Truong-Huy D Nguyen, and Mutlu Mete. Hand gesture controlled drones: An open source library. In *International Conference on Data Intelligence and Security*, 2018.

Marlon Oliveira, Houssem Chatbri, Ylva Ferstl, Mohamed Farouk, Suzanne Little, Noel O'Connor, and A. Sutherland. A dataset for Irish sign language recognition. In *Proceedings of the Irish Machine Vision and Image Processing Conference*, 2017.

Burr Settles. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 2012.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 1948.

Artem Shelmanov, Vadim Liventsev, Danil Kireev, Nikita Khromov, Alexander Panchenko, Irina Fedulova, and Dmitry V. Dylov. Active learning with deep pre-trained models for sequence tagging of clinical and biomedical texts. In *IEEE International Conference on Bioinformatics and Biomedicine*, 2019.

Kittasil Silanon and Nikom Suvonvorn. Finger-spelling recognition system using fuzzy finger shape and hand appearance features. In *International Conference on Digital Information and Communication Technology and its Applications*, 2014.

Alex Tamkin, Dat Nguyen, Salil Deshpande, Jesse Mu, and Noah Goodman. Active learning helps pretrained models learn the intended task. *arXiv*, 2022.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv*, 2017.

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. Including signed languages in natural language processing. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2021.

Appendix A. Background

A.1 Sign language Recognition

Hand gesture recognition is essential for applications such as human-computer interaction, robot control, and sign language assistance systems (Natarajan et al., 2018; Silanon and Suvonvorn, 2014; Elakkiya, 2021). The purpose of sign language alphabet recognition is to detect labels from isolated or continuous Signs (Yin et al., 2021). Recognition of signs using computational models is a challenging problem because of a number of reasons. In some cases, one gesture can look very different from a different angle. Another challenge is the variations of how a sign is performed by different signers (Mercanoglu and Keles, 2020). Dinh et al. (2006) built ASL alphabet classifiers by a combination of 24 (only static) sign detectors with AdaBoost, then using threshold information given by each detector to suggest which gesture is the best match. Elakkiya and Selvamani (2015) used active learning to utilize the availability of the current training classifier for increasingly improving performance. They do not use certain acquisition functions to measure uncertainty but directly choose the sample which is not correctly classified by the current classifier.

A.2 Active learning and Transfer active learning

Active learning is a machine learning method in which a learning system can query a human expert interactively to label data with the desired output(Cohn et al., 1996). The main goal of active learning is that active learner performance is better than traditional supervised learning with the same labeled data. Therefore active learners can interactively pose queries during the training stage and are part of the human-in-the-loop paradigm. Transfer Learning–Based Active Learning is a method to take advantage transfer learning for better active learning. Some papers tried to prove that pre-trained models are better active learners. (Hao et al., 2021) (Shelmanov et al., 2019)

A.3 Experiment for higher resolution of data

To ensure resolution does not limit the knowledge that the pre-trained model can transfer, we resize data in all four corpora to 96×96 and redo the same experiments. Because the ASL dataset used for the 28*28 experiment has a resolution of exactly 28*28, it cannot be used for resolution 96*96. We use a new ASL dataset from Kaggle[2] with an original resolution of 400*400 for this experiment. Because the computation time grows rapidly with the resolution, we made some adjustments to save time. Firstly, we only repeated all experiments three times for transfer active learning. Secondly, because the Irish dataset is augmented with rotation, we sample 30% of the Irish dataset and use it for experiments. Figure 5 reports test accuracy for all pre-training configurations, including no pre-training and double-hand fingerspelling(Indian sign language). Similar to low-resolution results, the most obvious improvement is pre-train on GSL and fine-tuning on ISL. Figure 6 shows the accuracy per class(gap). Interestingly, we also observe improvements for non-sharing alphabets, e.g. 'D', 'Q', 'R' in ASL and 'P', 'Q', 'X' in ISL. This shows that pre-training and fine-tuning can also help the model classify some non-shared alphabets.

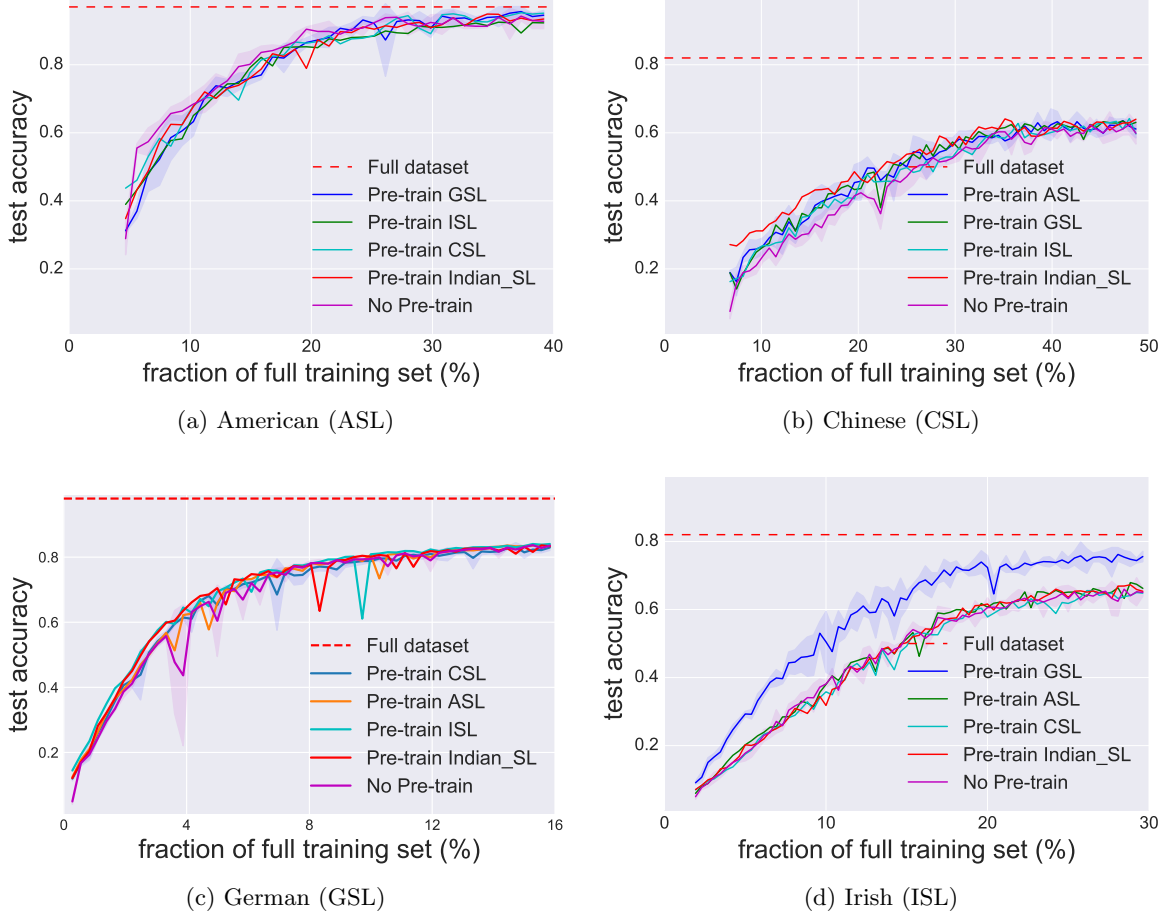2. https://www.kaggle.com/ayuraj/asl-dataset

Figure 5: *Transfer Active Learning*: The figure above shows AL results for each finger-spelling corpus with pre-training and no pre-training for resolution $= 96{\times}96$. Pre-training with a double hand-fingerspelling(Indian_SL) exists as a reference.
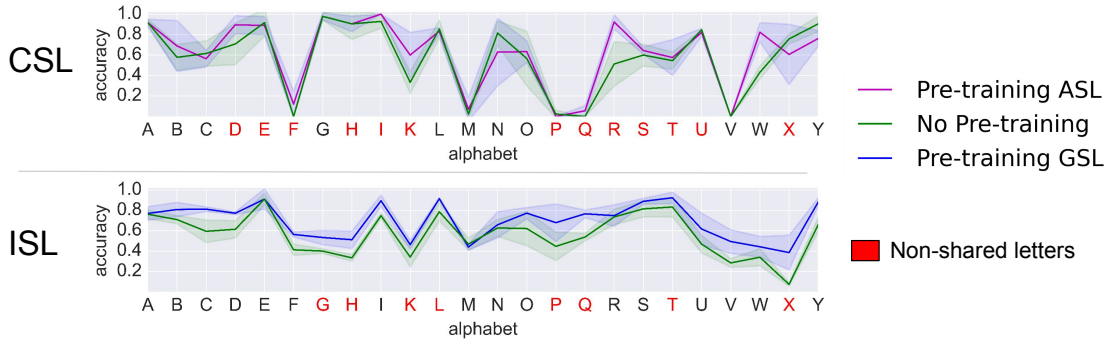

Figure 6: *Per Class Results for Transfer Active Learning at 96×96*: Generally, we see the gains from pre-training are had by shared letters (red) but not exclusively ('X' in ISL is not shared by GSL).