American Sign Language Recognition

Importance of Data Quality

Leonie Weiß
LT2326 Machine Learning for statistical NLP: Advanced
Gothenburg University
gusleoniwe@student.gu.se

Abstract

For machine learning tasks, reliable data is crucial. To gain some insight into the importance of data quality for machine learning tasks, this paper uses a 3D-Recurrent Neural Network to train and test a model for American Sign Language Recognition using a subset of the Word Level American Sign Language video dataset (WLASL). WLASL lacks linguistic precision in some of its labels and, therefore, creates the possibility to test to what extent the improvement of the labeled data impacts the model's performance in the testing. However, the resulting accuracies raise questions.

1 Introduction

Sign Language is very important for an inclusive community. However, a majority of the world's population lacks the needed knowledge to understand and use it. This can pose a big problem for the deaf and hard-of-hearing community. An automatic recognition of sign language can help this part of the population to communicate. A broader use of sign language in everyday life can help to empower the deaf community and motivate hearing people to learn sign language. A real-time recognition and translation application could for example be used for captioning virtual conferences like zoom meetings and would be a very helpful tool for deaf.

This project's aim is to automatically recognize American Sign language and to observe the impact of linguistic precision of the used data. To achieve this goal the project uses the publicly available Word-level American Sign Language Video Dataset (WLASL) and a 3D-Recurrent Neural Network to learn the features. The approach is taken from the paper "Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison" by Li et al. (2020). The data needed to recreate their project is published on GitHub[1] and, thereby, is easily accessible. Furthermore, this paper uses a spreadsheet of the ASLLRP Sign Bank to apply new labels to the videos contained by the WLASL and, hereby, improve its linguistic precision. Following this manipulation of the data, the model will be trained one more time and accuracies will be calculated anew. In a last step, those two models can be compared based on the calculated accuracies which should allow conclusions about the importance of the linguistic quality of the data used for training processes in machine learning to be drawn.

The first paragraph presents background information for the topic of sign language recognition and former work in the field. After that, the data and methods of this project will be explained. The following paragraph presents the results and the discussion of those results. Lastly, this report contains a conclusion that summarizes the project's results and proposes future work on the topic of ASL recognition and the work with the WLASL.

2 Sign Language Recognition

The topic of sign language recognition has already been extensively studied due to its relevance to a real-life problem. Wadhawan and Kumar (2021) find an increase in research papers on sign language recognition starting from 2012 while focusing on the years 2007 to 2017. This indicates an emerging trend of attention on sign language recognition and suggests a further increasing interest in the following years. Previous literature, however, differs in their approaches to the topic. Even though, a lot of work on sign language recognition has been done, no system has been

[1] See https://github.com/dxli94/WLASL/tree/master.

deployed on a large scale, so far, that can interpret a large vocabulary of signs in real-time.

There are different kinds of sign languages. Similar to spoken languages, their exist American, Indian, Arabic, German, etc. sign language. Most of the found literature focuses on American Sign Language (cf. e.g. Lee et al., 2021; Li et al., 2020; Sharma & Kumar, 2021). Since American Sign Language (ASL) is also the predominately used sing language, with an estimated count of users in a range from 2,500,000 to 5,000,000 (Wadhawan & Kumar, 2021: 790) from more than 20 different countries (Li et al., 2020: 1), this paper chose to work with ASL.

The literature also presents different kinds of approaches in terms of language level. Sign language can use letters to sign something (cf. Abdulhussein & Raheem, 2020). However, just like in spoken language this is mostly used for spelling out a word. The remaining two levels are words and sentences. Word-level sign language recognition is also known as isolated sign language recognition because it focuses on isolated signs for certain words, while sentence-level sign language recognition is known as continuous sign language recognition looking at continuous sequences of signed words in a row. This project focuses on words because this sequence is more approachable for a first try of recognizing sign language. The recognition of signed sentences should work in a very similar way. However, the differentiating of the different words would pose an additional problem for an automatic recognition.

Additionally, in some projects not only hand and arm gestures but also facial expressions and body postures are observed (cf. Wadhawan and Kumar, 2021). Those are non-manual elements which are used for some signs. However, for manual signs only hands and arms are used.

Lastly, different articles work with different data types. Some researchers use images (cf. Abdulhussein and Raheem, 2020; Wadhawan & Kumar, 2020) and, thereby, focus on static sign language while others use videos and focus on dynamic sign language. Others use leap motion controllers (cf. Lee et al.) as a source to collect data that should be classified in the next step. This project will focus on video recognition to test the possible practicability of a real-life tool working with dynamic video data that should be captioned with the signed sentences.

The method used for the specific sign language recognition problem is influenced by the different approaches. Abdulhussein and Raheem (2020), for example, use a Convolutional Neural Network (CNN) to recognize static ASL letters using a dataset of images representing an alphabet of 24 letters for ASL and a softmax layer for classification. They chose 10 images per letter and reach an accuracy of 99.3%. Wadhawan and Kumar (2020) also use a CNN to train a model using 350 images for each of 200 distinct signs to recognize Indian Sign Language. They not only use letters but also 10 digits and 67 commonly used Indian words. Using CNNs for static sign language recognition seems to be very common. Wadhawan and Kumar (2021) found that in the years 2007-2017 the majority of projects on ASL uses neural networks (33%), support vector machines (21%), hybrid techniques (21%) or CNNs (13%).

2.1 Challenges of Sign Language Recognition

Sign language is, just like any other language, very complex. The meaning of the signs is the result from a combination of body motions, manual movements and head poses. Even subtle differences can lead to different meanings. Those differences have to be recognized by a recognition method to be used with a high accuracy. Additionally, this is complicated by the enormous size of the vocabulary of sign languages which exceeds the size of categories used in similar gesture recognition tasks by more than a few hundreds or even thousands (Li et al., 2020: 1-2). Furthermore, the meaning of signs can be context-dependent. For example, some nouns and verbs with the same lemma are sometimes represented by the same sign. To react to those difficulties of sign language recognition, Li et al. (2020) built a dataset of great size to make a training for SLR models possible, while achieving high accuracies.

3 Data and Methods

The data used for this project is the Word-Level American Sign Language dataset (WLASL). It was created by the authors Li et al. (2020) of the paper “Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison” that has already been mentioned in the introduction and preceding paragraph. The dataset can be downloaded on

Kaggle[2] or created by following the instructions of Li's GitHub repository[3] containing the project. The WLASL contains videos for 2,000 different words. The videos are taken from multiple educational sign language websites, like ASLU[4] and ASL-LEX[5], as well as ASL tutorial videos on YouTube. From YouTube Li et al. (2020) chose videos clearly naming the signed word in the title. Each word has a minimum amount of seven different instances in the video corpus. This makes it possible to use this dataset for training and testing. Furthermore, Li et al. (2020: 4) argue that less occurrences of certain signs in the educational sign languages websites used to compile the data imply that the word is not commonly used in everyday life. Therefore, it is not of disadvantage for the model trained on the WLASL and its future usage to delete those videos.

Before the selection procedure used to balance the dataset, the WLASL contained 34,404 video samples of 3,126 ASL glosses from 20 different websites (Li et al., 2020: 4). Every gloss chosen for the WLASL is composed by only one sign and every single sign containing more than one English word was removed to make sure that the WLASL only contains words. The instances are signed by different signers and filmed in a nearly-frontal view in front of different backgrounds. Additionally to the gloss label for each video, the WLASL contains some meta data. Firstly, the signer is annotated. Diversity of the signers is important to train the model for a real-world use case. Secondly, the temporal boundary is annotated containing the start and end frame of the sign in the video. Thirdly, the data contains information about the body bounding.box. To identify the signer and crop the box to its size, Li et al. (2020) used YOLOv3, which is a person detection tool. Lastly, the data contains a dialect variation annotation. Using manually annotation and a set of dialect labels, the WLASL can be balanced in dialect variation. The authors distributed the labeled data in training and testing data and removed all sign variations with less than five examples to ensure that training, validation and testing is possible. However, these variations are usually not commonly used in everyday life (Li et al., 2020: 5).

The WLASL dataset can be used in four different subsets using the top-K most common glosses in the dataset with K = {100, 300, 1000, 2000}. The subsets are called WLASL100, WLASL3000, WLASL1000 and WLASL2000. To make the training feasible for this project, the smallest subset WLASL100 was chosen.

Li et al. (2020) compare different methods in their paper. The one chosen for this paper is the use of a 3D Convolutional Neural Network because this achieved the best results in the comparison. The 3D-CNN is used to extract the spatial and temporal information of the frames of the videos. Li et al. use pretrained I3D weights by Carreira et al. (2017). The I3D model is trained on the image database *ImageNet* and finetuned on the dataset *Kinetics-400*, that contains videos of human movements. This model was finetuned with the Adam optimizer using the WLASL. The final layer of the 3D Convolutional Neural Network is a softmax layer to classify the videos. Training is stopped after 216 epochs containing 64,000 training steps in total. To evaluate the model Li et al. use a top-K classification accuracy with K = {1, 5, 10}. They argue that the classification will be easier for longer video sequences with more signs forming a sentence because of the given and additionally learnable context. This project will look at those accuracies to get a more general answer to the question, how feasible automatic sign language recognition is. The training and testing process was done twice: once with the original WLASL100 subset and once after the improvement of the labels of the WLASL100. A following comparison of the achieved accuracies makes it possible to observe changes in the results due to the improvement of the data quality.

3.1 Dataset Improvement

The WLASL dataset contains a few errors in the labels of the data. Neidle and Ballard (2022) state that there is no "1-1 relationship" (ibid., 3) between English glosses and ASL signs and, additionally, no standardized convention for association of English glosses and American sign language signs. Therefore, taking the English glosses mentioned in the titles of the videos found on the internet as labels for the WLASL leads to inaccuracies in the labeling of the data. Similar to spoken language, there is variability in meaning when

[2] See https://www.kaggle.com/datasets/risangbaskoro/wlasl-processed.

[3] See https://github.com/dxli94/WLASL/tree/master.

[4] See https://www.lifeprint.com/.

[5] See https://asl-lex.org/.

using ASL signs. The WLASL has problems with instances that are represented by the same ASL sign but have different English glosses as well as instances that are the same English gloss but have different ASL signs. For example, “close” as the opposite of “to open” and in the sense of “near” are represented by two different ASL signs but appear for the same gloss in the WLASL (Neidle & Ballard, 2022: 3).

Since only part of the data of the WLASL is taken from educational sign language websites, the errors probably occur in the data scraped from YouTube. The data quality of the educational SL websites is guaranteed due to checks by experts before uploading the data. This guarantee for the quality of the data is not given for YouTube videos because in this case no such quality checks are required to upload the data. Furthermore, the individual video collections from the different sources that are combined to build the WLASL do not use the same consistent label system (Neidle & Ballard, 2022: 2). Therefore, the combination of those different publicly shared ASL video corpora leads to an inconsistency of the WLASL’s annotation.

The American Sign Language Linguistic Research Project (ASLLRP) offers a table with corrected labels for the WLASL[6]. They use the labels of a label set that they created for their own dataset of sign language videos, the ASLLVD[7]. Neidle and Ballard (2022) have established this label set to ensure a 1-1 correspondence between the English gloss labels and the ASL signs. The project described in this report uses the WLASL with corrected labels to see if there is a difference in the results after improving the linguistic quality of the data.

To improve the WLASL100 dataset the labels were changed in the word class list provided by Li et al. (2020) as a txt-file containing the labels and the label IDs. In the following step, the json-file combining video IDs and labels was changed and used to create the new dataset WLASL122. Before selection, the dataset WLASL100 contained 100 different labels. Following the application of the labels proposed by Neidle and Ballard (2022) to the WLASL100, the labels and videos that do not have at least seven instances where excluded. This ensures that the amount of data is large enough to be divided in training and testing set. After the application of the new labels, the videos covered 169 English glosses containing 47 labels that did not occur at least seven times. Therefore, new dataset contains 122 labels.

The dataset with the new labels was named by the pattern used by Li et al. (2020) for their different sizes of the WLASL. It will be called WLASL122 in the following parts of the paper because it contains videos for 122 different glosses and, therefore, covers 122 labels.

After the improvement of the dataset described in this section of the report, the training was done in the same way again on this new dataset WLASL122.

4 Results

	top-1	top-5	top-10
WLASL100	0.738	0.9058	0.9358
WLASL122	0.4078	0.6878	0.7657

Table 1: Top-1, top-5, top-10 accuracy achieved after training with different subsets WLASL100 and WLASL122.

The accuracies of the two different models show a big difference and can be seen in table 1. The results of the training on WLASL122 show a top-1 accuracy of about 40.78%, being lower than an accuracy by chance, while the training on WLASL100 achieves a top-1 accuracy of 73.8%. Even the top-10 accuracies differ a lot being 76.57% for WLASL122 and 93.58% for WLASL100.

The accuracies might be lower for the new dataset with revised labels because, contrary to the original dataset, the new dataset is not as balanced regarding dialect. On a contrary, Neidle and Ballard (2022) state that the dialect annotations are “highly problematic” (ibid., 3), since some of the videos mistakenly annotated with the same gloss because of sharing the lexical form of the word, even though, having different meanings, like close with two different meanings, have been annotated as the same word with dialectal variants. This simply is wrong. The accuracy of the dialectal variant annotation is not fully determined.

Furthermore, the new label set counts more different labels. A bigger group of labels to choose from leads to lower accuracies. This can also be seen in the results of Li et al. (2020: 7). They report the top-1 accuracy (%) for the WLASL100

[6] See https://dai.cs.rutgers.edu/dai/s/aboutwlasl.

[7] See https://dai.cs.rutgers.edu/dai/s/index;jsessionid=E5326918820FCC6FAD287BC77315A01F?redirect=signbankdownload.

being 65.89, for the WLASL300 being 56.14, for the WLASL1000 being 47.33 and for the WLASL2000 being 32.48. The WLASL122 resulting in a lower accuracy follows this pattern. However, it is lower than the results achieved by Li et al. using the WLASL1000, so there has to be another reason.

Another reason for the low accuracy achieved with the WLASL122 could be a lower number of videos for each label since the amount of labels grew but the amount of videos remained the same. Less videos that can be used for training lead to worse results.

However, the results raise questions about the reason for those low accuracies, since the raised number of labels and smaller amount of videos per label do not fully explain the low accuracies.

5 Conclusion

The theoretical part of this report proves the importance of data quality. However, this proof could not be replicated in the results of the project using the WLASL100 and WLASL122. The low accuracy of 40.78% achieved by using the WLASL122 and the big difference to the accuracy of 73.8% achieved by using the WLASL100 can be explained by the smaller number of videos per English gloss and by more labels being used. However, those reasons do not satisfyingly solve the question about the big difference between the accuracies remaining. To find more reasons for this big of a difference the amount of videos for each label could be checked.

To further explore this topic, the already changed WLASL corpus could be supplemented by videos from the ASLLVD corpus. In this way, all words from the WLASL100 can be used because the added videos would ensure that every word is represented by enough videos to do a training, evaluation and testing.

References

Abdulwahab A. Abdulhussein, and Firas A. Raheem. 2020. Hand gesture recognition of static letters American sign language (ASL) using deep learning. *Engineering and Technology Journal, 38 (06),* 926-937. https://doi.org/10.30684/etj.v38i6A.533.

Joao Carreira, and Andrew Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. *Computer Vision and Pattern Recognition.* https://doi.org/10.48550/arXiv.1705.07750.

C.K.M. Lee, Kam K.H. Ng, Chun-Hsien Chen, H.C.W. Lau, S.Y. Chung, and Tiffany Tsoi. 2021. American sing language recognition and training method with recurrent neural network. In *Expert Systems with Applications 167*. https://doi.org/10.1016/j.eswa.2020.114403.

Dongxu Li, Cristian Rodriguez Opazo, Xin Yu, and Hongdong Li. 2020. Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison. https://arxiv.org/abs/1910.11006.

Carol Neidle, and Carey M. Ballard. 2022. Why Alternative Gloss Labels Will Increase the Value of the WLASL Dataset. https://www.researchgate.net/publication/362541595

Shikhar Sharma, and Krishan Kumar. 2021. ASL-3DCNN: American sign language recognition technique using 3-D convolutional neural networks. In *Multimedia Tools and Applications 80,* 26319-26331. DOI: 10.1007/s11042-021-10768-5 .

Ankita Wadhawan, and Parteek Kumar. 2020. Deep learning-based sign language recognition system for statistic signs. In *Neural Computing and Applications 32,* 7957-7968. https://doi.org/10.1007/s00521-019-04691-y.

Ankita Wadhawan, and Parteek Kumar. 2021. Sign Language Recognition Systems: A Decade Systematic Literature Review. In *Archives of Computational Methods in Engineering 28*, 785-813. https://doi.org/10.1007/s11831-019-09384-2.

A Supplementary Material

The following list shows the glosses from WLASL100 the original model was trained on.

0	book
1	drink
2	computer
3	before
4	chair
5	go
6	clothes
7	who
8	candy
9	cousin
10	deaf
11	fine
12	help
13	no
14	thin
15	walk
16	year
17	yes

18	all
19	black
20	cool
21	finish
22	hot
23	like
24	many
25	mother
26	now
27	orange
28	table
29	thanksgiving
30	what
31	woman
32	bed
33	blue
34	bowling
35	can
36	dog
37	family
38	fish
39	graduate
40	hat
41	hearing
42	kiss
43	language
44	later
45	man
46	shirt
47	study
48	tall
49	white
50	wrong
51	accident
52	apple
53	bird
54	change
55	color
56	corn
57	cow
58	dance
59	dark
60	doctor
61	eat
62	enjoy
63	forget
64	give
65	last
66	meet
67	pink
68	pizza
69	play
70	school
71	secretary
72	short
73	time
74	want
75	work
76	africa
77	basketball
78	birthday
79	brown
80	but
81	cheat
82	city
83	cook
84	decide
85	full
86	how
87	jacket
88	letter
89	medicine
90	need
91	paint
92	paper
93	pull
94	purple
95	right
96	same
97	son
98	tell
99	thursday

The following list contains the labels of the WLASL122 used for the second training.

0	BASKETBALL
1	ORANGE
2	CITY/COMMUNITY
3	VOLUNTEER/SHIRT
4	SHIRT
5	WHO
6	BUT
7	PLAY
8	BIRTHDAY
9	MAN
10	MANY
11	SAME
12	TELL
13	HOW
14	ADULT-TALL
15	TALL
16	BIRD
17	COLOR

18	COAT
19	YES
20	DOCTOR
21	DOG
22	PURPLE
23	PAST
24	BEFORE
25	BLACK
26	WHITE
27	PINK
28	BROWN
29	DRINK
30	WANT
31	TIME
32	WALK
33	DANCE
34	HEARING
35	KISS
36	SON
37	DARK
38	NOW
39	NS_AFRICA
40	GIVE
41	COUSIN
42	APPLE
43	FORGET
44	BOOK
45	HAT
46	CORRECT
47	HELP
48	PAPER
49	THINK+DECIDE
50	#NO
51	WHAT
52	LATER_2
53	ALL
54	#ALL
55	GO
56	SCHOOL
57	DEAF
58	FINISH
59	ALSO
60	CHEAT
61	MOTHER
62	STUDY
63	FINE
64	NEAT
65	PULL
66	WOMAN
67	EAT
68	LANGUAGE
69	SIT
70	ENJOY
71	PAINT
72	KISS_3
73	BLUE
74	BED
75	MEDICINE
76	LAST
77	FAMILY
78	FISH
79	“WHAT”
80	(1)CHEAT
81	WRONG
82	SHORT
83	CHAIR
84	RIGHT
85	COOK/KITCHEN
86	COMPUTER
87	DRESS/CLOTHES
88	THURSDAY
89	SECRETARY
90	GRADUATE
91	(G)THANKSGIVING
92	THANKSGIVING
93	WORK
94	PIZZA_4
95	PIZZA
96	THIN
97	SLIM
98	FULL
99	CAN
100	MISTAKE
101	CUTE
102	CHANGE
103	WHO_2
104	CANDY
105	TABLE
106	BOWLING
107	YEAR
108	COMPUTER_6
109	COMPUTER_5
110	MEET
111	CLOTHES
112	HOT
113	LETTER/MAIL
114	CORN
115	SHOULD
116	FED-UP/FULL
117	LIKE
118	COW
119	SHORT-HEIGHT
120	LATER
121	ACCIDENT