Real Time Sign Language Recognition

Karthik S K

Mid Thesis Report

MAY 2022

**Abstract**

Sign languages are visual communication techniques used mostly by deaf people all over the world. Waving fingers, hands, arms, and facial gestures are the most frequent ways to sign words and sentences. Sign languages are fully formed languages with their own syntax and lexicon. Sign language recognition is a complex task. Several factors are employed to identify signs, including hand form and orientation, hand movements, body posture, and facial expressions. Even with cutting-edge models, tackling this problem computationally for a large vocabulary of signs in real-world circumstances remains a difficulty. Indian Sign Language (ISL) is a complete language with its own grammar, syntax, lexicon, and other distinguishing linguistic features. It is used by about 5 million deaf people in India. There is presently no publicly available dataset on ISL to test Sign Language Recognition techniques (SLR). In this research study, The Indian Lexicon Sign Language Dataset (INCLUDE) is considered for study. The dataset contains 0.27 million frames from 4,287 films and 263 different word signs from 15 different word categories. INCLUDE is documented with the help of expert signers to achieve a close resemblance to natural conditions. Numerous deep neural networks incorporating diverse approaches for augmentation, feature extraction, encoding, and decoding have been assessed. The dataset's size and quality make it possible to test deep models for Sign Language Recognition on ISL. A comparison of different deep learning models is shown, and a model with good accuracy is identified. The most accurate model has a 94.5 percent accuracy rate. This model just trains a decoder and employs a pre-trained feature extractor and encoder. It improves on previous results and provides a quick way to support SLR in various languages.

**Table of Contents**

Abstract 2

[LIST OF FIGURES 4](#_Toc102512208)

[LIST OF TABLES 4](#_Toc102512209)

[LIST OF ABBREVIATIONS 4](#_Toc102512210)

[1. Background 5](#_Toc102512211)

[2. Related Research 6](#_Toc102512212)

[3. Research Questions (If any) 10](#_Toc102512213)

[4. Aim and Objectives 10](#_Toc102512214)

[5. Significance of the Study 10](#_Toc102512215)

[6. Scope of the Study 11](#_Toc102512216)

[7. Research Methodology 11](#_Toc102512217)

[8. Required Resources 20](#_Toc102512218)

[9. Research Plan 20](#_Toc102512219)

[References 20](#_Toc102512220)

# LIST OF FIGURES

Figure 1: The training process for two stream I3D on Kinetics dataset 11

Figure 2: CNN architecture with Softmax output 12

Figure 3: VGG-16 Architecture 13

Figure 4: Confusion Matrix 19

# LIST OF TABLES

Table 1: INCLUDE: Category Size 9

Table 2: INCLUDE: Key statistics 10

# LIST OF ABBREVIATIONS

|  |  |
| --- | --- |
| SL | Sign Language |
| SLR | Sign Language Recognition |
| SLT | Sign Language Translation |
| CSLR | Continuous Sign Language Recognition |
| ISL | Indian Sign Language |
| EDA | Exploratory Data Analysis |
| CNN | Convolutional Neural Network |
| VGG | Visual Geometry Group |
| NMT | Neural Machine Translation |
| I3D | Inflated 3D |
| HMM | Hidden Markov Model |
| RNN | Recurrent Neural Network |
| LSTM | Long Short-Term Memory |
| WASL | Word-level American Sign Language |
| AUTSL | Ankara University Turkish Sign Language |
| GRU | Gated Recurrent Units |

# 1. Background

The amount of video content produced and consumed is steadily increasing. As a result, deep learning is increasingly being used for tasks involving video data, such as video categorization, object detection, object tracking, action recognition, visual question answering, and video encoding. Many of these tasks have practical applications, and some of them have significant accessibility implications. Sign language recognition (SLR) can be regarded of as an action recognition task (Yao et al., 2015) . Sign languages are visual communication techniques that are predominantly utilized by deaf people all over the world. Words and phrases are most signed by waving with fingers, hands, arms, and facial expressions. Sign languages have their own syntax and lexicon and are fully formed languages. Furthermore, they differ from place to region and are frequently incomprehensible to one another, despite some similarities. In terms of lexicon and articulation rate, they also differ from the spoken languages of a given region (Huang et al., 2018). Sign languages, as the deaf community's principal means of communication, have their own linguistic frameworks. The goal of sign language interpretation systems is to automatically translate sign languages using techniques such as vision. Word-level sign language recognition (also known as "isolated sign language recognition") and sentence-level sign language recognition (also known as "continuous sign language recognition") are the two fundamental tasks in this procedure.

Sign language recognition using computer models is a difficult problem for a variety of reasons. To begin, fine-grained analysis of the local and global motion of numerous body components, such as the hand, arms, and face, is required. Although the hand motions for some pairs of signs appear to be extremely identical, the distinctions in face expressions distinguish the meaning. Depending on the number of repetitions, a very similar hand gesture can take on a distinct meaning. Another problem is the differences in how different signers perform a sign, such as body and position variances, duration variations of different parts of the signals, and so on. Variation in illumination and backdrop complicates the challenge, which is a problem in computer vision. When the number of signs in the corpus grows, these difficulties become more difficult to solve (Sincan and Keles, 2020).

Indian Sign Language (ISL) is a complete language with its own grammar, syntax, lexicon, and other linguistic characteristics. ISL differs from its western equivalents in several ways, the most notable of which being its lexicon, which has a high level of iconicity. While most other sign languages feature a few compound signs (signs that are made up of two or more signs), ISL has a widespread compounding system. Compounding signals for Male and Sibling, for example, are used to sign the word Brother, whereas compound signs for Female and Marry are used to sign the term Wife. This is also a deviation from Indian spoken languages, which are noted for their large number of intricate kinship phrases. The variety of possible compositions allows ISL to have a very wide lexicon. Another way in which ISL differs from other sign languages is in the amount of space required for signing around the body. In ISL, the top signing space represents distance and authority, whereas in many European and North American Sign Languages, distance is represented by a horizontal plane in front of the signer. These and other differences between ISL and other sign languages need the development and testing of novel methods for ISL Sign Language Recognition. Furthermore, India's vast population of deaf individuals – over 5 million – emphasizes the significant practical significance of any deployable solution (Sridhar et al., 2020).

The availability of resources in the form of big, standardized datasets is critical for the development and evaluation of machine learning models. Importantly, none of these datasets are publicly accessible, preventing additional research. These datasets also have two significant flaws. To begin with, the number of classes, or different word signs, is extremely minimal. The videos are frequently limited in some way. Only the hands are visible in the datasets because the photographs were cropped. Pictures or films with consistent backdrops, with the signer's clothing generally matching the background color. The usefulness of models trained on the dataset in real-world scenarios is limited by these dataset restrictions. As a result, no ISL dataset of sufficient size and quality for machine learning research is currently available.

The first publicly available ISL dataset, INCLUDE - Indian Lexicon Sign Language Dataset, has been proposed to address the absence of public datasets and scalable algorithms for ISL recognition. Data augmentation, feature extraction with pre-trained networks, and encoding and decoding were all part of the deep learning workflows for SLR on the INCLUDE dataset. The feature extraction was merged with a pre-trained pose detection network, and the encoding was done with a pre-trained MobileNet network, and the decoding was done with trained bidirectional LSTMs.

# 2. Related Research

The research survey related to solving the problem of Sign Language Recognition (SLR) has been carried out and is summarized in this section. The hearing-impaired community around the world use some variants of sign languages for communication and hence it is meaningful and important for automatic translation of a sign language. The problems that are prevalent in the domain of SLR are mainly two - isolated SLR and continuous SLR (Huang et al., 2018). Isolated SLR recognizes word by word, whereas continuous SLR translates entire sentences.

The existing methods for continuous SLR while using isolated SLR as building blocks also add layers of pre-processing and post-processing which indicate temporal segmentation and sentence synthesis respectively. Most existing SLRs fall into the category of isolated SLR which deals with the recognition of words or expressions. Continuous SLR being more challenging involves reconstructing sentence structures, which divide sentence-to-sentence recognition problem into three stages - temporal segmentation of videos, isolated word/expression recognition, and sentence synthesis with a language model.

Most approaches in the recent past show great progress in SLR which are based on the Hidden Markov Model (HMM) with various features such as motion trajectory. The transitional movements between signs which are not correctly captured are addressed using the approaches of offline training and online recognition. The threshold matrix and rate thresholds are proposed in offline training, where each element of the matrix indicates the minimum probability when a segment belongs to a sign, and rate thresholds are defined as average probability for signs. If the evaluation of a certain segment is smaller than all the thresholds, then it is regarded as a transitional movement which should be removed. In the online recognition stage, Dynamic Time Warping (DTW) and Length-Root method is used to record the time intervals for fine segmentation and the endpoint for each candidate sign is determined (Zhang et al., 2014).

The temporal segmentation even though popular is difficult to implement due to below reasons:

* The transitional movement between hand gestures can be subtle and ambiguous.
* Inaccurate segmentation can incur significant performance penalty on subsequent steps.
* The isolated SLR step that deals with the recognition of words or expression requires per-video-frame labels which are extremely time consuming.

A novel end-to-end sequence-to-sequence model that can generate captions for videos is a relevant research area. Recurrent Neural Networks, specifically LSTMs which demonstrated state-of-the-art performance in image caption generation, is trained on video-sentence pairs and the model learns associating a sequence of video frames to sequence of words to describe an event in the video clip (Venugopalan et al., 2015). Attention mechanisms can also be incorporated into LSTM for automatic selection of most likely video frames (Yao et al., 2015).

Various research groups and industry labs have developed open-source toolkits for Neural Machine Translation (NMT). These are based on open-source deep learning platforms. However, these tools are targeted only towards research groups with a solid background in machine translation and deep learning, along with experience in navigating, understanding, and handling large code bases. With all these, none of the existing NMT tools has been designed primarily for readability or accessibility for novices, which makes it challenging to understand how NMT is implemented, the features each toolkit implements and what toolkit can be chosen to code their own project as fast and simple as possible.

Joey NMT (Kreutzer et al., 2019) designed for novices provides clean, well-documented, minimum code which is of comparable quality to more complex codebases on standard benchmarks. The core features include standard network architectures (RNN, transformer, different attention mechanisms, input feeding, configurable encoder/decoder bridge), standard learning techniques (dropout, learning rate scheduling, weight tying, early stopping criteria), and visualization/monitoring tools.

Sign languages are under-resources with corpora recorded typically for linguistic research and this does not provide type/token ratios needed for statistical natural language processing. This data significantly differs from real language encountered outside the research lab. RWTH-PHOENIX-Weather 2012 corpus for German sign language (DGS) recorded and annotated real life sign language footage aired in 2009 and 2010 by the German public TV station “PHOENIX” in the context of weather forecasts. This corpus has been significantly extended by adding more data with bilingual annotation in DGS glosses and written German, creating the RWTH-PHOENIX-Weather 2014 corpus. There are also annotations for spatial positions of the hands and face of a signer for over 40,000 video frames along with annotations for hand shapes and orientations on the frame level (Schmidt et al., 2014).

The methods that were employed previously for continuous sign language recognition involved hidden Markov models (HMM) with limited capacity for capturing temporal information. With deep neural networks, a framework for continuous sign language recognition transcribes videos of sign language sentences to sequences of ordered gloss labels. Deep Convolutional Neural Networks (CNN) with stacked temporal fusion layers for feature extraction modules, and bidirectional Recurrent Neural Networks (RNN) for sequence learning modules were proposed with an iterative optimization process to exploit the representational capability of deep neural networks with limited data. The end-to-end recognition model for alignment proposal was trained first, and then the proposal was used to tune the feature extraction module. This process can run iteratively to achieve improvements on the recognition performance (Cui et al., 2019).

The prior studies in Sign Language Translation (SLT) have shown that effectively recognizing individual signs improves the performance of translation to a great extent. A novel transformer architecture jointly learns Continuous Sign Language Recognition (CSLR) and Translation while being trainable in an end-to-end manner. The recognition and translation problems are addressed in a single unified architecture called Connectionist Temporal Classification (CTC) loss (Camgoz et al., 2020). The contributions can be summarized as:

* The novel formalization of Continuous Sign Language Recognition (CSLR) and Sign Language Translation (SLT) which is multi-task uses the power of glosses without limiting the translation to spoken language.
* The state-of-the-art results achieved in successful application of transformers for Continuous Sign Language Recognition (CSLR) and Sign Language Translation (SLT), outperforming all comparable previous approaches.
* New baseline results to guide future research in this field.

Indian Sign Language (ISL) which is used by over 5 million deaf people in India is a complete sign language with unique linguistic attributes with one grammar, syntax, and vocabulary. To carry out and evaluate Sign Language Recognition (SLR) approaches, currently there is no publicly available dataset on ISL. Earlier efforts in creating datasets for ISL include video datasets and image datasets with a single image denoting a sign. Since none of these datasets are in the public domain, any further study is a challenge. Other limitations are with the number of classes i.e., the number of distinct word signs being significantly low, and images being cropped so that only certain portions (only hands) are visible in the datasets. These constraints on the dataset pose challenges in limiting the models that can be built and trained for real world scenarios. The size and quality expected for Machine Learning (ML) research are not available for ISL dataset.

The lack of public datasets and scalable models for ISL recognition has been addressed by proposing the first publicly available ISL dataset called INCLUDE - Indian Lexicon Sign Language Dataset. The dataset contains 0.27 million frames with 263 classes from 15 different word categories. There are videos recorded with the help of deaf community, with the background, resolution and lighting chosen to resemble real world scenarios. A smaller subset named INCLUDE-50 with 50 classes was also proposed for rapid evaluation. The deep learning pipelines for SLR on INCLUDE dataset included data augmentation, feature extraction with pre-trained networks, using encoding and decoding. The feature extraction combined with pre-trained pose detection network, encoding with pre-trained MobileNet network, and decoding with trained bidirectional LSTMs has an accuracy of 94.5% on INCLUDE-50 and 85.6% on INCLUDE (Sridhar et al., 2020).

Deaf persons can use vision-based sign language recognition to communicate with others. Most available sign language databases, on the other hand, are confined to a small number of words. Models learned from those datasets cannot be used due to the low vocabulary size. A new large-scale Word-Level American Sign Language (WLASL) video dataset with over 2000 words performed by over 100 signers was proposed (Li et al., 2019) . This dataset will be made available to the research community on the public domain. It is the largest publicly available ASL dataset for word-level sign recognition research. Different deep learning approaches for word-level sign identification were tested and evaluated their performance in large-scale scenarios using this new large-scale dataset. Two different models were used and compared: (i) a holistic visual appearance-based approach and (ii) a 2D human pose-based approach. Both models provide valuable baselines for method benchmarking in the community. Furthermore, the study offers a unique pose-based temporal graph convolution networks (Posture-TGCN) that jointly models spatial and temporal relationships in human pose trajectories, which has improved the pose-based method's performance. Findings reveal that pose-based and appearance-based models perform similarly on 2,000 words/glosses, with top-10 accuracy up to 62.63 percent, showing the dataset's validity and difficulties.

Sign language identification is a difficult challenge in which signs are identified by numerous sources of simultaneous local and global articulations, such as hand shape and orientation, hand movements, body position, and facial expressions. Even with state-of-the-art models, solving this problem computationally for a wide vocabulary of signs in real-world scenarios remains a challenge. A new large-scale multi-modal Turkish Sign Language dataset (AUTSL) with a benchmark and performance evaluation baseline models was proposed in this paper (Sincan and Keles, 2020) . In total, there are 38,336 isolated sign video samples in our collection, which includes 226 signs performed by 43 individual signers. The samples include a wide range of backgrounds captured in both indoor and outdoor settings. Furthermore, throughout the recordings, signers' spatial placements and postures change. Each sample was captured using the Microsoft Kinect v2 and includes RGB color, depth, and skeletal modalities. For user-independent model evaluations, the benchmark training and test sets were created. Employed Convolutional Neural Networks (CNNs) to extract features and unidirectional and bidirectional Long Short-Term Memory (LSTM) models to characterize temporal information, and provided empirical evaluations using the benchmark. To boost the performance of our models, added feature pooling modules and temporal attention. On the AUTSL and Montalbano datasets, tested the baseline models.

On the Montalbano dataset, the models achieved competitive results with state-of-the-art approaches, achieving 96.11 percent accuracy. The models achieved 95.95 percent accuracy in AUTSL random train-test splits. The best baseline model achieved 62.02 percent accuracy in the proposed user-independent benchmark dataset. The disparities in the results of the identical baseline models demonstrate the difficulties in the benchmark dataset.

# 3. Research Questions (If any)

The following research questions are suggested for each of the research objectives as highlighted as follows.

# 4. Aim and Objectives

The main aim of this research is to propose a ………………………………. The identification of the breast cancer ………….

The research objectives are formulated based on the aim of this study which are as follows:

* To analyze the pattern and relationship between the risk factors ………….
* To suggest a suitable balancing technique …………..
* To compare between the predictive models ……………..
* To evaluate the performance of ……………….

# 5. Significance of the Study

Sign languages are visual languages that use hand, face, and body motions to communicate. Learning a second language is not only beneficial to your brain, but it can also help you enhance your communication abilities. Early man utilized signs to communicate, according to some scholars, long before spoken language was established. Even though today's world progressed significantly, sign languages in their most basic form are still used such as pushing index finger between lips to calm a noisy child, raising hand to hail a cab, or pointing to an item on the menu.

Sign language is the link that connects us to the world of persons who are deaf or have difficulty speaking. A variety of hand, finger, arm, head, and face expressions are used to help the deaf with those around them and vice versa. It enables individuals to comprehend the world around them through visual descriptions and hence contribute to society. Some of the benefits of sign language include - assists persons who are deaf or hard of hearing aiding social inclusions. It provides deaf children an opportunity to educate themselves. Non-deaf volunteers who volunteer to learn sign language in order to communicate with the disabled are instilled with a sense of social responsibility as well as sensitivity.

# 6. Scope of the Study

# 7. Research Methodology

Sign language is a system of communication using visual gestures and signs as used by the deaf community and is the main mode of communication. Deaf and people who have problems hearing use sign language between their own community and other people. Computer recognition of sign languages covers sign gesture recognition and continues till text/speech generation. Sign gestures can be classified as static and dynamic with static gesture recognition being simpler than dynamic gesture recognition. While most of the approaches in recent times show progress in Sign Language Recognition (SLR) which are based on Hidden Markov Model (HMM), temporal segmentation techniques are also used to record time intervals for fine segmentation and the endpoint for each candidate sign is determined. Generating captions for videos using novel end-to-end sequence-to-sequence models is a relevant research area. Attention mechanisms are used in Recurrent Neural Networks (RNN), specifically LSTMs. In this research paper, as part of research methodology the following sections are covered which includes target dataset description, data pre-processing, different models and algorithms and evaluation metrics.

**7.1 Dataset Description**

Indian Lexicon Sign Language Dataset - INCLUDE - an ISL (Indian Sign Language) dataset contains 0.27 million frames across 4287 videos over 263-word signs from 15 different word categories. This dataset is recorded with the help of experienced signers. Each video is a recording of 1 ISL sign, signed by deaf students from St. Louis School for the Deaf, Adyar, Chennai (Sridhar et al., 2020).

Based on the 2 main principles - The videos resembling real life scenarios, and a dataset covering a diverse set of signs with multiple videos, the dataset was created by recording with the help 7 experienced signers from the school for the deaf. Each class is a sign in the dataset, which is signed by multiple signers, containing 2 to 6 videos per class.

The classes belong to 15 broad word categories which are popular words in ISL. This is shown in Table 1.1. The category-wise summary statistics are shown in Table 1.2.

Table 1: INCLUDE: Category Size

|  |  |  |
| --- | --- | --- |
| **Category** | **Number of Classes** | **Number of Videos** |
| Adjectives | 59 | 791 |
| Animals | 8 | 166 |
| Clothes | 10 | 198 |
| Colours | 11 | 222 |
| Days and Time | 22 | 306 |
| Electronics | 10 | 140 |
| Greetings | 9 | 185 |
| Means of Transport | 9 | 186 |
| Objects at Home | 27 | 379 |
| Occupations | 16 | 225 |
| People | 26 | 513 |
| Places | 19 | 399 |
| Pronouns | 8 | 168 |
| Seasons | 6 | 85 |
| Society | 23 | 324 |
| **Total** | **263** | **4287** |

Table 2: INCLUDE: Key statistics

|  |  |
| --- | --- |
| **Characteristic** | **INCLUDE** |
| Categories | 15 |
| Words | 263 |
| Videos | 4287 |
| Avg. Videos per Class | 16.3 |
| Avg. Video Length | 2.57s |
| Min. Video Length | 1.28s |
| Max. Video Length | 6.16s |
| Frame Rate | 25fps |
| Resolution | 1920x1080 |

**7.2 Data Preprocessing**

The preprocessing techniques on video data include - VGG (Visual Geometry Group) Feature Extraction, and I3D (Inflated 3D ConvNet) Feature Extraction & Vision Transformer. In this section each of the above feature extraction techniques are analyzed.

The temporal element is one of the key differences between information in a single image and information in a video. As a result, deep learning model architectures have been improved to include 3D processing to process temporal data as well. Through the I3D model, this article summarizes the architectural changes from photos to video. Researchers from DeepMind and the University of Oxford presented the I3D model in a paper (Carreira and Zisserman, 2017). The research examines prior approaches to the problem of video action detection while also offering a novel architecture, which is the emphasis of this paper. Their method starts with a two-dimensional design and then inflates all the filters and pooling kernels. They offer an additional dimension to be considered by inflating them, which in our instance is time. Filters in 2D models are square N x N, but when they are inflated, they become cubic N x N x N.

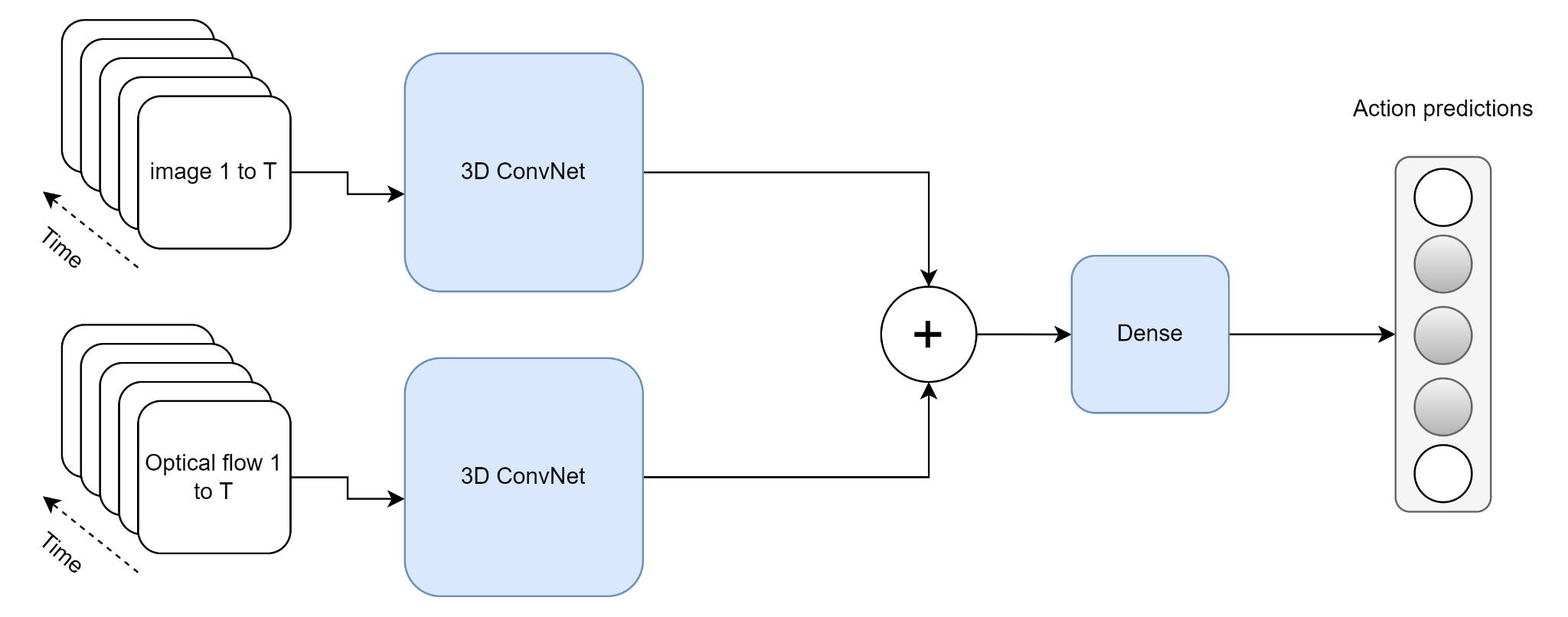


Figure 1: The training process for two stream I3D on Kinetics dataset

Since the dataset contains videos, some of the pre-processing techniques such as imputing missing values, outlier treatment is not applicable, and hence these are not mentioned in this section.

**7.3 Model Building**

In this section different models and algorithms that will be used for effective Sign Language Recognition are explained. The following methods are used:

* Method 1: 3D CNN based Classification with Softmax in the last layer
* Method 2: VGG Feature Extraction + 2D CNN/NN based Classification
* Method 3: Method 1/Method 2 with or without Attention Mechanism
* Method 4: VGG Features + Transformers Encoder + GRU based Decoder
* Method 5: I3D Pretrained/Video Vision Transformer + Transformers Encoder + GRU based Decoder

**7.3.1 Method 1: 3D CNN based Classification with Softmax in the last layer**

Convolutional Neural Networks (CNN) are used in a variety of applications. It is, without a doubt, the most widely used deep learning architecture. The enormous popularity and effectiveness of convnets has sparked a recent rise in interest in deep learning. AlexNet sparked interest in CNN in 2012, and it has grown rapidly since then. Researchers went from an 8-layer AlexNet to a 152-layer ResNet in just three years. CNN has become the go-to model for any image-related issue. They outperform the competitors in terms of accuracy. The fundamental advantage of CNN over its predecessors is that it discovers essential traits without the need for human intervention. In addition, CNN is computationally efficient. It performs parameter sharing and uses special convolution and pooling algorithms. CNN models can run on any device, making them globally appealing (Gu et al., 2015)

Fig 2 shows the architecture of CNN with Softmax output in the last layer.

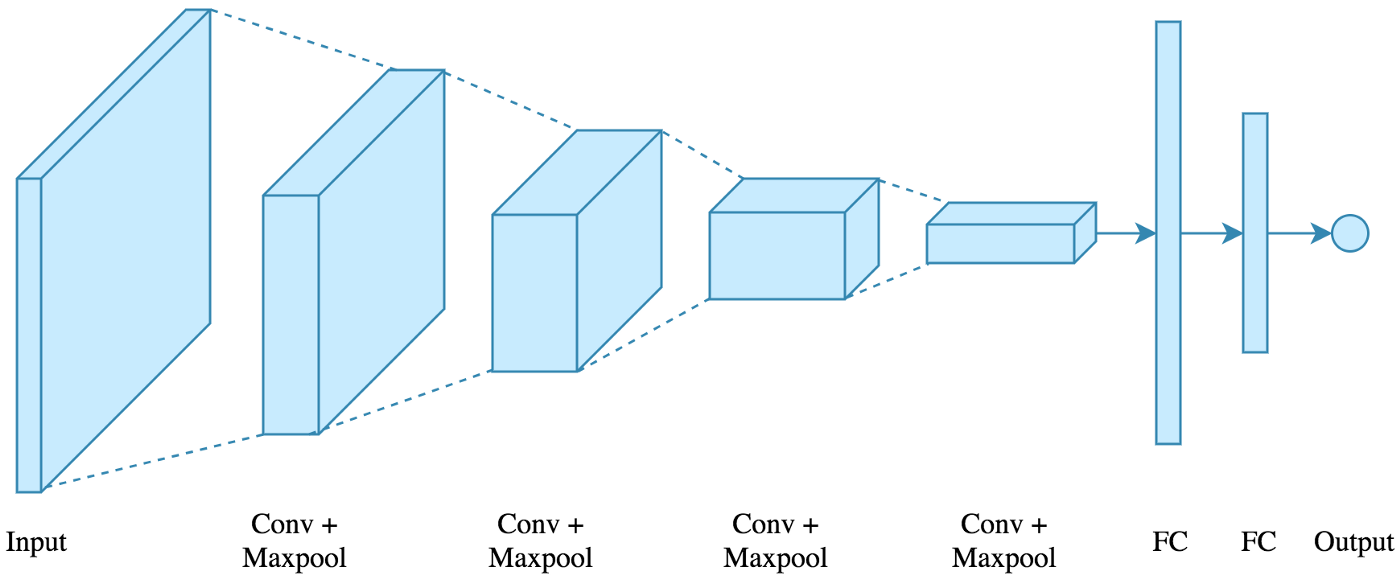


Figure 2: CNN architecture with Softmax output

The following architecture will be used: 4 convolutions + pooling layers will be followed by 2 fully connected layers. The input is an image, and the output is multi class. There are 4 methods which are used and are indicated below:

* Conv2D: A convolutional layer is created using this method. The filter count is the first parameter, while the filter size is the second. As an activation function, relu non-linearity is used with Padding enabled. There are 2 options for Padding: same or valid. Same indicate padding with the number on the edge, while valid indicates no padding. For convolution layers, the stride is set to 1 by default.
* MaxPooling2D: The only argument is the window size, which is used to create a maxpooling layer. Because it is the most frequent, a 2x2 window is used. The default stride length is retained, which is equal to the window size 2.
* Flatten: Flatten the output of the convolution + pooling layers before feeding it into the fully linked layers
* Dropout: Dropout is a simple concept that is used to prevent overfitting. During training, a neuron is momentarily "dropped" or inhibited with probability p at each repetition. This signifies that at this iteration, all this neuron's inputs and outputs will be disabled. At each training step, the dropped-out neurons are resampled with probability p, so a dropped-out neuron at one step can become active at the next. The dropout-rate hyperparameter p is commonly a number around 0.5, which corresponds to 50 percent of the neurons being dropped out.

**7.3.2: Method 2: VGG Feature Extraction + 2D CNN based Classification**

VGG16 is a CNN (Convolutional Neural Network) that is widely regarded as one of the best computer vision models available today. The creators of this model analyzed the networks and enhanced the depth using an architecture with very small (3x3) convolution filters, which outperformed previous-art setups significantly. The depth was increased to 16–19 weight layers, resulting in 138 trainable parameters. VGG16 is a 92.7 percent accurate object identification and classification system that can classify 1000 images into 1000 different categories. It's a common picture classification algorithm that's simple to utilize with transfer learning (Simonyan and Zisserman, 2014)

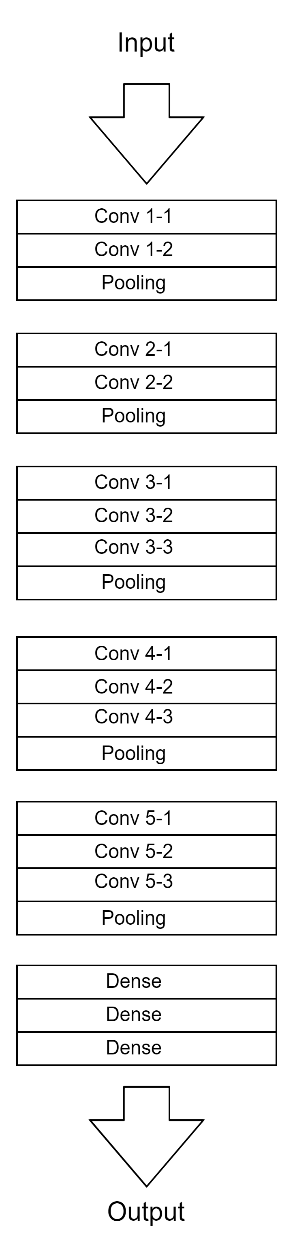


Figure 3: VGG-16 Architecture

Figure 3 shows the VGG -16 architecture. The key pointers to note:

* The 16 in VGG16 stands for 16 weighted layers. VGG16 comprises thirteen convolutional layers, five Max Pooling layers, and three Dense layers, for a total of twenty-one layers, but only sixteen weight layers, or learnable parameters layers.
* VGG16 uses a 224, 244 input tensor size with 3 RGB channels.
* The most distinctive feature of VGG16 is that, rather than having a huge number of hyper-parameters, it uses 3x3 filter convolution layers with stride 1 and always used the same padding and maxpooling layer of 2x2 filter stride 2.
* The convolution and maxpooling layers are placed in a regular pattern throughout the architecture.
* Conv-1 layer has 64 filters, Conv-2 layer has 128 filters, Conv-3 layer has 256 filters, and Conv 4 and Conv 5 layers have 512 filters.
* Following a stack of convolutional layers, three Fully Connected (FC) layers are added: the first two have 4096 channels each, while the third performs 1000-way image classification and so has 1000 channels (one for each class). The soft-max layer is the final layer.

Challenges:

* It takes a longer time to train (the initial VGG model took 2–3 weeks to train on the Nvidia Titan GPU).
* VGG-16 trained weights are 528 MB in size. As a result, it consumes a significant amount of storage space and bandwidth, making it inefficient.

**7.3.3 Method 3: Method 1/Method 2 with Attention mechanism**

This is the study (Bahdanau et al., 2014) that established the now-famous "Attention Mechanism". Even though the concept of attention has evolved, the mechanism described in this study is still recognized as "Bahdanau Attention". The study describes the following:

* The notion of employing Neural Networks to translate phrases from a source language to a target language is known as Neural Machine Translation (NMT). Until this study, such NMT models have relied on numerous networks, each of which had to be trained separately.
* The research proposes that a single, massive neural network be built and trained to comprehend a sentence and correctly translate it, which is the foundation for all current Sequence to Sequence models based on Encoder-Decoder architecture.
* Machine Translation is analogous to finding a target sentence y that maximizes the conditional probability of p(y|x), where x is the source sentence, from a probabilistic standpoint.
* The goal of an NMT task is to use a Parallel training corpus to maximize the Conditional Probability of Sentence Pairs. To simulate such a relationship, a parameterized model would be employed, with Backpropagation utilized to learn the parameter weights.

A source sentence is fed into an encoder, which converts it into a fixed-length vector. The translation (target sentence) from the Encoded Vector is output by a Decoder. For a given source-target sentence pair, the Encoder-Decoder system is jointly trained to maximize the conditional probability of an accurate translation. There are some limitations with encoder-decoder architecture. For information about the source sentence, the Decoder only uses the last encoded fixed-length vector. It's very difficult for the Encoder to compress all the information into a single vector when the source sentence is quite long. The performance of a basic encoder-decoder degrades significantly as the length of a source sentence increases, according to actual evidence.

The research proposes an Encoder-Decoder model extension that learns to 'align' and 'translate' together. When the NMT model generates a translated term, it does soft search for a set of positions in the source sentence and looks for the positions with the highest concentration of relevant information. It's like selecting the words that make the most sense in the final translation. This is incompatible with the idea of storing the full source sentence into a single fixed-length context vector. The NMT model then predicts a target translation using context vectors associated with these source positions as well as previously generated translation outputs. The source text is encoded as a sequence of vectors, and the decoder selects a subset of these vectors to produce the translation. It allows the NMT model to interpret long words and do a selective search based on context importance rather than squashing all the information into a single vector.

**7.3.4 Method 4: VGG Features + Transformer’s Encoder + GRU based Decoder**

Due to advances in Sequence Modelling, such as the comeback of Long-Short Term Memory networks (LSTMs) and the development of Gated Recurrent Units (GRUs), video captioning and summarization have been highly popular in recent years. Existing architectures use CNNs to extract spatio-temporal information and soft attention layers to model dependencies using GRUs or LSTMs. These attention layers aid in paying to the most conspicuous aspects and improve the recurrent units; nonetheless, these models suffer from the recurrent units' intrinsic flaws. With the introduction of the Transformer model, the Sequence Modeling field has taken a new turn. 3D CNN architectures like C3D and Two-stream I3D for video extraction in this study [citation] is used to construct a Transformer-based model for video captioning. Certain dimensionality reduction techniques are used to keep the model's total size under control.

**7.3.5 Method 5: I3D Pretrained/Video Vision Transformer + Transformer’s Encoder + GRU based Decoder**

Rather than employing frame-level feature extractors, networks to extract spatio-temporal information from videos are used directly. 3D convolutions are used in these structures to encode both spatial and temporal information in videos. Using 2D convolutions on an image or a video (series of frames) results in a single feature map. Using 3D convolutions on a set of frames, on the other hand, produces a set of feature maps. The size of the temporal kernel and the strides employed determine the number of feature mappings.

Recent advancements in the field of activity recognition have resulted in a variety of designs that can be used to extract spatio-temporal features. Instead of depending on a recurrent network to encode information from each time step, architectures that can directly offer temporal information are looked at. Features are extracted for the Transformer model using I3D (Inflated 3D) Convolutional Neural Networks for Activity Recognition (Bilkhu et al., 2019)

**7.4 Evaluation Metrics**

Accuracy is one of the metrics to describe the accuracy of an algorithm on a classification task. Since the dataset is balanced, Accuracy as the evaluation metric is used to measure the performance of the models. It is the number of samples that are paired divided by number of samples.

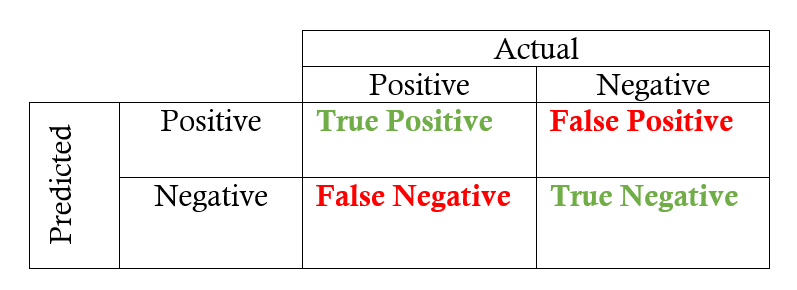


Figure 4: Confusion Matrix

Figure 4 shows the confusion matrix. Accuracy is the proportion of true results among the total number of cases examined.

Accuracy = (True Positive + True Negative) / (True Positive + False Positive + False Negative + True Negative)

Accuracy is a valid choice of evaluation for classification problems which are well balanced and not skewed or No class imbalance.

Top-1 Accuracy is the conventional accuracy, model prediction (the one with the highest probability) must be exactly the expected answer. It measures the proportion of examples for which the predicted label matches the single target label.

Top-5 Accuracy means any of the model’s top 5 highest probability answers matches with the expected answer. It considers a classification correct if any of the 5 predictions matches the target label.

# 8. Required Resources

The research will need below hardware and software resources throughout the implementation.

**8.1 Software Requirements**

Operating System: Ubuntu/Mac OS/Windows

* Programming Language: Python 3.9.1, Shell Script
* Package Manager: pip
  + Python Libraries:
    - OpenCV
    - NLTK
    - Matplotlib
    - Numpy
    - CSV

#### Hardware Requirements

A laptop with below configuration will be used:

* SSD: 512GB
* Ram: 40GB (depends on batch size)
* Graphics: NVIDIA 2080 RTI, 12GB

# 9. Research Plan

# References

Bahdanau, D., Cho, K. and Bengio, Y., (2014) Neural Machine Translation by Jointly Learning to Align and Translate. [online] Available at: http://arxiv.org/abs/1409.0473.

Bilkhu, M., Wang, S. and Dobhal, T., (2019) Attention is all you need for Videos: Self-attention based Video Summarization using Universal Transformers. [online] Available at: http://arxiv.org/abs/1906.02792.

Camgoz, N.C., Koller, O., Hadfield, S. and Bowden, R., (2020) Sign Language Transformers: Joint End-to-end Sign Language Recognition and Translation. [online] Available at: http://arxiv.org/abs/2003.13830.

Carreira, J. and Zisserman, A., (2017) Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. [online] Available at: http://arxiv.org/abs/1705.07750.

Cui, R., Liu, H. and Zhang, C., (2019) A Deep Neural Framework for Continuous Sign Language Recognition by Iterative Training. *IEEE Transactions on Multimedia*, 217, pp.1880–1891.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, L., Wang, G., Cai, J. and Chen, T., (2015) Recent Advances in Convolutional Neural Networks. [online] Available at: http://arxiv.org/abs/1512.07108.

Huang, J., Zhou, W., Zhang, Q., Li, H. and Li, W., (2018) Video-based Sign Language Recognition without Temporal Segmentation. [online] Available at: http://arxiv.org/abs/1801.10111.

Kreutzer, J., Bastings, J. and Riezler, S., (2019) Joey NMT: A Minimalist NMT Toolkit for Novices. [online] Available at: http://arxiv.org/abs/1907.12484.

Li, D., Opazo, C.R., Yu, X. and Li, H., (2019) Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison. [online] Available at: http://arxiv.org/abs/1910.11006.

Schmidt, C., Koller, O., Bellgardt, M., Forster, J. and Ney, H., (2014) *Extensions of the Sign Language Recognition and Translation Corpus RWTH-PHOENIX-Weather*. [online] Available at: www.signspeak.eu.

Simonyan, K. and Zisserman, A., (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition. [online] Available at: http://arxiv.org/abs/1409.1556.

Sincan, O.M. and Keles, H.Y., (2020) AUTSL: A Large Scale Multi-modal Turkish Sign Language Dataset and Baseline Methods. [online] Available at: http://arxiv.org/abs/2008.00932.

Sridhar, A., Ganesan, R.G., Kumar, P. and Khapra, M., (2020) INCLUDE: A Large Scale Dataset for Indian Sign Language Recognition. In: *MM 2020 - Proceedings of the 28th ACM International Conference on Multimedia*. Association for Computing Machinery, Inc, pp.1366–1375.

Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T. and Saenko, K., (2015) Sequence to Sequence -- Video to Text. [online] Available at: http://arxiv.org/abs/1505.00487.

Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H. and Courville, A., (2015) Describing Videos by Exploiting Temporal Structure. [online] Available at: http://arxiv.org/abs/1502.08029.

Zhang, J., Zhou, W. and Li, H., (2014) A Threshold-based HMM-DTW Approach for Continuous Sign Language Recognition. In: *Proceedings of International Conference on Internet Multimedia Computing and Service - ICIMCS ’14*. [online] New York, New York, USA: ACM Press, pp.237–240. Available at: http://dl.acm.org/citation.cfm?doid=2632856.2632931.