PANIMALAR ENGINEERING COLLEGE

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CS8811 PROJECT WORK

REVIEW NO:

Sign Language Recognition, Gloss Text, and Translation to Spoken Language: Utility for Aurally Challenged.

Guide Name: **Dr. Senthil Kumar G**

Batch No: **E1**

Team Members with Register number:

Isaac G (211418104091)

Jebastin Stephen Sarjin John (211418104096)

Project Details.

PROJECT TITLE	**Sign Language Recognition, Gloss Text, and Translation to Spoken Language: Utility for Aurally Challenged.**	
DOMAIN	**NLP, Deep Learning, Neural Machine Transformer, Sign Language Translation.**	
TEAM MEMBERS	**Isaac G (211418104091)**	**Jebastin Stephen Sarjin John (211418104098)**
PROJECT GUIDE	**Dr. Senthil Kumar G**	

Introduction.

Technology has aided today's communication very drastically. There are various means namely text, images, videos and other media by which information is shared and thereby communication is facilitated.

But a problem exists:

(1). Persons who are challenged with aural disorders and hearing disabilities find it difficult to understand audio information.

(2). Since there are diverse languages in the world, it would be difficult to translate one language to a language known to the particular person.

So in order to aid them, deep learning contributes features such as **Continuous Sign Language Recognition** (CSLR).

There would be some possible sign representations while communicating (i.e., hand shape, facial expression and body posture), Such signs has to be captured, analyzed and then converted to gloss language, which would be an intermediate between the Sign representation and the Spoken language.

We use two modules:

The Spatial-Modeling Cue (SMC) module is concerned with spatial representation and categorizes visual features of different cues with the aid of a self-contained pose estimation branch.

The Temporal Modeling Cue (TMC) module models temporal correlations along two parallel paths, i.e., intra-cue and inter- cue, which aims to preserve the uniqueness and explore the collaboration of multiple cues.

Finally, we design a joint optimization strategy to achieve the end-to-end sequence learning of the STMC network.

Literature Survey.

S.NO	Journal Type With Year	Authors	Title
1.	arXiv:2202.10419	Kayo Yin; Graham Neubig.	Interpreting Language Models with Contrastive Explanations.
2.	arXiv:2105.07476	Amit Moryossef; Kayo Yin; Graham Neubig; Yoav Goldberg.	Data Augmentation for Sign Language Gloss Translation.
3.	arXiv:2004.00588	Kayo Yin; Jesse Read.	Better Sign Language Translation with STMC-Transformer.
4.	arXiv:1409.0473	Bahanau, D.; Cho, K.; and Bengio, Y. 2014.	Neural Machine Translation by jointly Learning to Align and Translate in ICLR.

Problem statement.

- This project is aimed for helping normal persons understand sign language with much ease.
- And also help the aurally challenged persons to understand the sign language.
- *eg*: Lets take a person A who is aurally challenged, to communicate he must either know sign language or he can't communicate. So to help the person A in both scenarios this project is developed. Through which sign language may be converted to spoken language to help the normal people understand, so that communication can take place.
- *eg:* Lets take person B who is normal person and he wants to either understand sign language or communicate to a deaf person, This project will help them understand what the aurally challenged person is saying.

Hardware/Software used.

SOFTWARE SPECIFICATION:

Technology Used: NLP, Deep Learning, OpenNMT (Neural Machine translation), CSLR (Continuous Sign Language Recognition).

Operating System: Windows 7 & above / Ubuntu

Software Requirement: Python 3.8, Anaconda

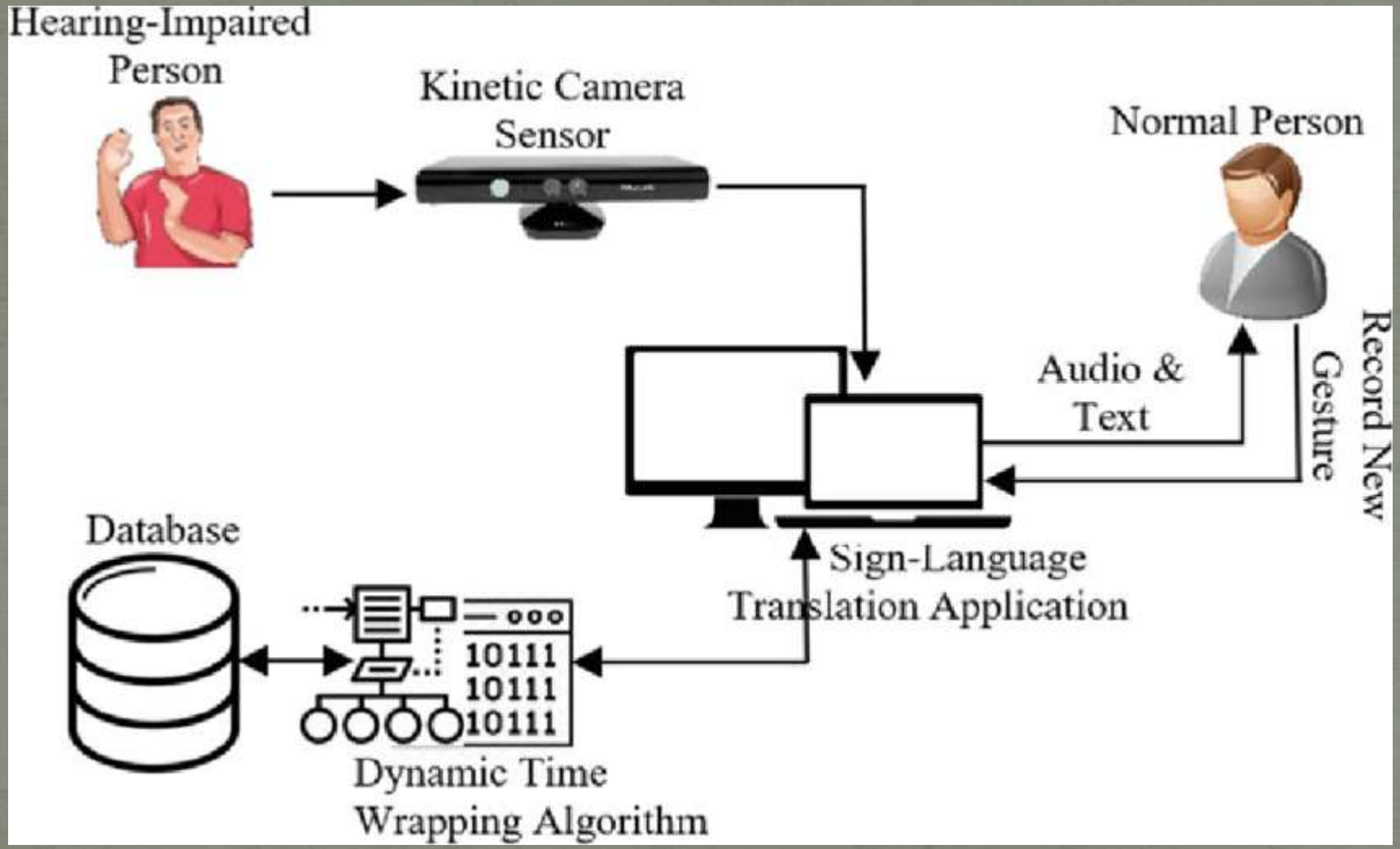
SYSTEM SPECIFICATIONS:

Ram: 8GB, 64-bit OS.

Processor: i3 6^{th} Gen+ Intel Processor.

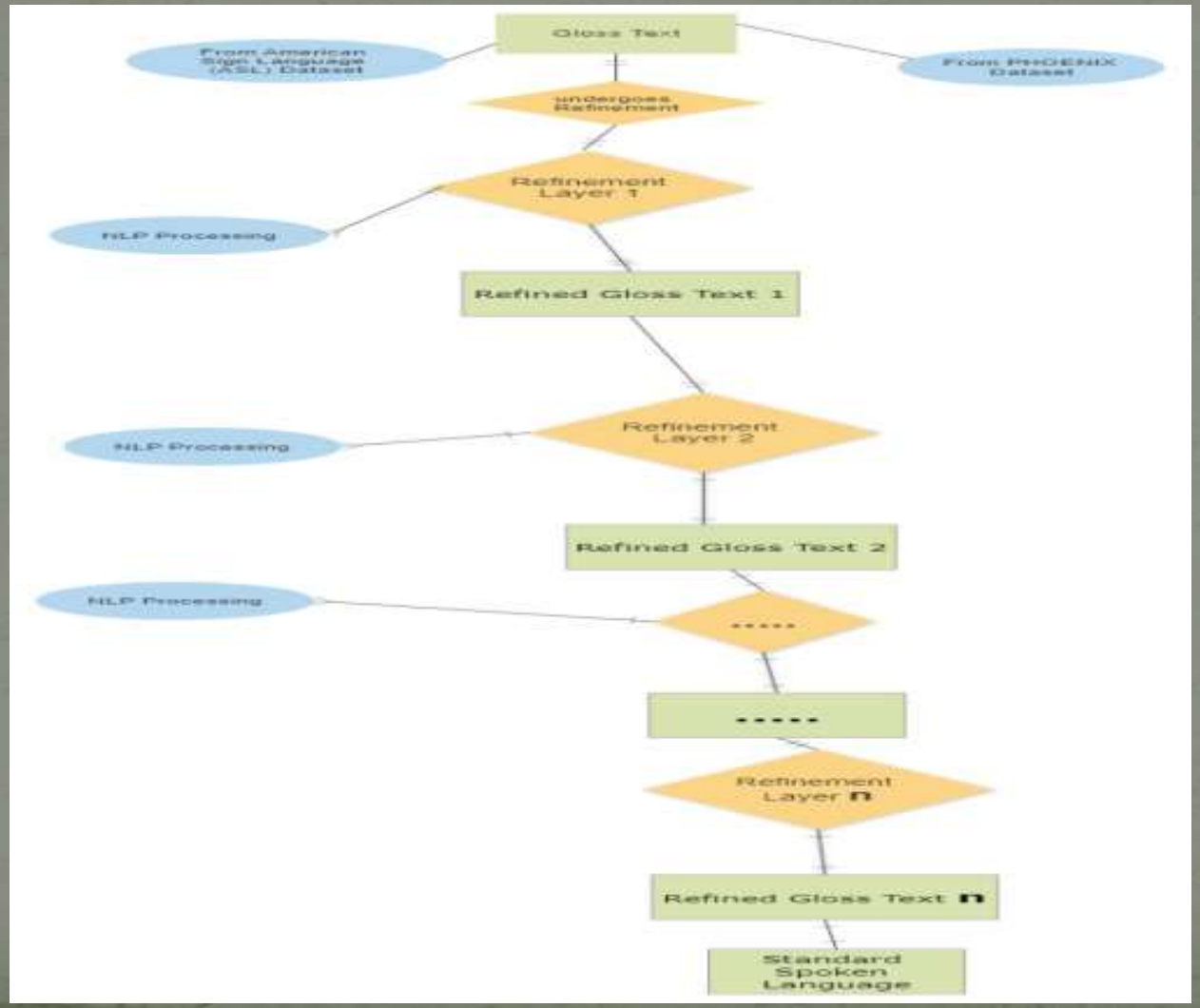
Hard Disk Capacity: 128 GB+

System Architecture.

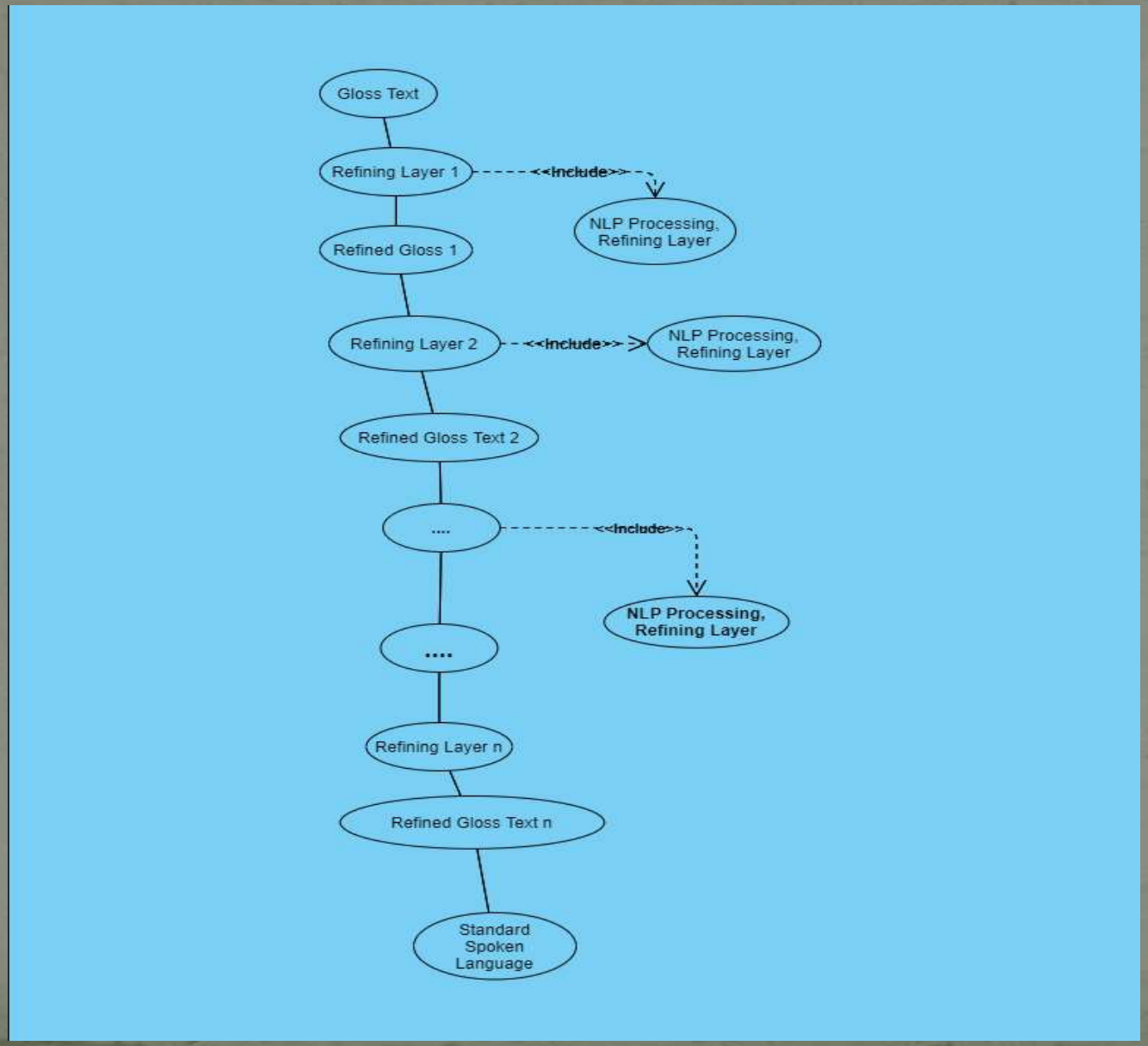


System Design.

(1.) ER Diagram:



(2.)Use-Case Diagram:



Modules.

(1.) Functionalities:

- **BLSTM Encoder:**

We use BLSTM Encoder to process the gloss text and the working of the encoder is given below;

With the help of BLSTM, The state transitions in the sequence of inputs has to be modeled using internal state of recurrent neural networks (RNN). Sign gloss sequence has to be mapped with spatial-temporal feature sequence using RNN.

RNN takes the feature sequence as input & then generates T^1 hidden state as:

$$\mathbf{h_t = RNN(h_{t-1}, 0_t)}$$

So, with the help of BLSTM Encoder, The Gloss text is processed.

- **Datasets:**

We Experiment the module with the help of 3 datasets:

1. PHOENIX-2014
2. CSL
3. PHOENIX-2014-T

The Data sets can be used to train, test, and validate the model.

(1.) PHOENIX-2014-T:

It is an extended version of PHOENIX-2014 and has two-stage annotations for new videos. One is sign gloss annotations for new videos. Another is German translation annotations for sign language translation (SLT) task. The videos are split into counts of 7096, 519 and 642 for Train, Dev and Test respectively. It has not overlap with the previous version between Train, Dev and Test set. The vocabulary size is 1115 for sign gloss and 300 for German.

(2.)CSL:

It is a Chinese Sign Language dataset, which has 100 sign language sentences about daily life with 178 words. Each sentence is performed by 50 signers and there are 5000 videos in total. For pre-training, it also provides a matched isolated Chinese Sign Language database, which contains 500 words. Each word is performed 10 times by 50 signers.

(3.)PHOENIX-2014:

It is a publicly available German Sign Language dataset, which is the most popular benchmark for CSLR. The corpus was recorded from broadcast new about the weather. It contains videos of 9 different signers with a vocabulary size of 1295. Videos are split into counts of 5672, 540 and 629 for Train, Dev and Test respectively. Our method is evaluated on the multi-signer database.

- **CSLR:**

  Computer Supported Learning Resource is function of performing continuous sign recognition in Real-time, Signs are sensed with CSLR.
  In this project to perform experiments we use datasets of PHOENIX-2014, CSL, PHOENIX-2014-T.
  CSLR is a task that recognizes the grammar of sign language itself, and it is difficult to provide meaningful interpretation only with CSLR because the grammar of sign language and the grammar of the spoken language are different.

- **Spatial multi-cue module:**

  The SMC module is dedicated to spatial representation and explicitly decomposes visual features of different cues with the aid of a self-contained pose estimation branch.
  In SMC module we perform 3 actions: pose estimation, Patch cropping, Feature generation.
  It consider the spatial movements or angles of a particular action.

- **Temporal Multi-cue model:**

The Temporal Multi-cue (TMC) module models temporal correlations along two parallel paths, i.e., intra-cue and inter-cue, which aims to preserve the uniqueness and explore the collaboration of multiple cues. Finally, we design a joint optimization strategy to achieve the end-to-end sequence learning of the STMC network.
Temporal multi-cue (TMC) module intends to integrate spatiotemporal information from two aspects, intra-cue and inter-cue. The intra-cue path captures the unique features of each visual cue. The inter-cue path learns the combination of fused features from different cues at different time scales.
The TMC module contains two paths:

(1.) Intra-cue path:

Unique features of different cues at different time scales are provided. The temporal transformation inside each cue is performed as follows;

$$f_{l,n} = \mathrm{ReLU}(\mathcal{K}_k^{\frac{G}{N}}(f_{l-1,n})),$$
$$f_l = [f_{l,1}, f_{l,2}, \cdots, f_{l,N}].$$

(2.) Inter-cue path:

The temporal transformation has to be performed on the inter-cue feature from the previous block and fuse information from the intra-cue path as follows;

$$o_l = \mathrm{ReLU}([\mathcal{K}_k^{\frac{C}{2}}(o_{l-1}), \mathcal{K}_1^{\frac{C}{2}}(f_l)]),$$

After each block, a temporal max-pooling with stride *2* and Kernel size *2* is performed.We use two blocks in the TMC module.

- **Spatial Multi-cue module:**

The SMC module is dedicated to spatial representation and explicitly decomposes visual features of different cues with the aid of a self-contained pose estimation branch.
It consider the spatial movements or angles of a particular action.
In SMC module we perform 3 actions:

- pose estimation
- Patch cropping
- Feature generation

(1.)Pose Estimation:

Deconvolutional networks are widely used in pixel-wise prediction. For pose estimation, Two deconvolutional layers of VGG-11. The stride of each layer is 2. So, the feature maps are *4x* upsampled from the resolution *14x14* to *56x56*. The output is fed into a point-wise convolutional layer to generate *K* predicted heat maps. In each heat map, the position of its corresponding keypoint is expected to show the highest response value.

(2.) Patch cropping:

In CSLR, the perception of detailed visual cues is vital, including eye gaze, facial expression, mouth shape, hand shape and orientations of hands. Our model takes predicted positions of the nose and both wrists as the center points of the face and both hands. The patches are cropped from the output of 4^{th} convolutional layer of VGG-11. The cropping sizes are fixed to *24x24* for both hands and *16x16* for the face. It is large enough to cover body parts of a signer whose upper body is visible to the camera. The cropping sizes are fixed to *24x24* for both hands and *16x16* for the face. It is large enough to cover body parts of a signer whose upper body is visible to the camera. The center point of each patch is clamped into a range to ensure that the patch would not cross the border of the original feature map.

(3.) Feature Generation:

After K keypoints are predicted, They are flattened to a 1D-vectorwith dimension *2K* and passed through two fully-connected (FC) layers with ReLU to get the feature vector of pose cue. Then, Feature maps of the face and both hands are cropped and processed by several convolutional layers, separately. Most sign gestures rely on the cooperation of the both hands. So we use weight-sharing convolutional layers for both hands. The outputs of them are concatenated along the channel-dimension. Finally, we perform global average pooling over all the feature maps with spatial dimension to form feature vectors of different cues.

Inference Time:

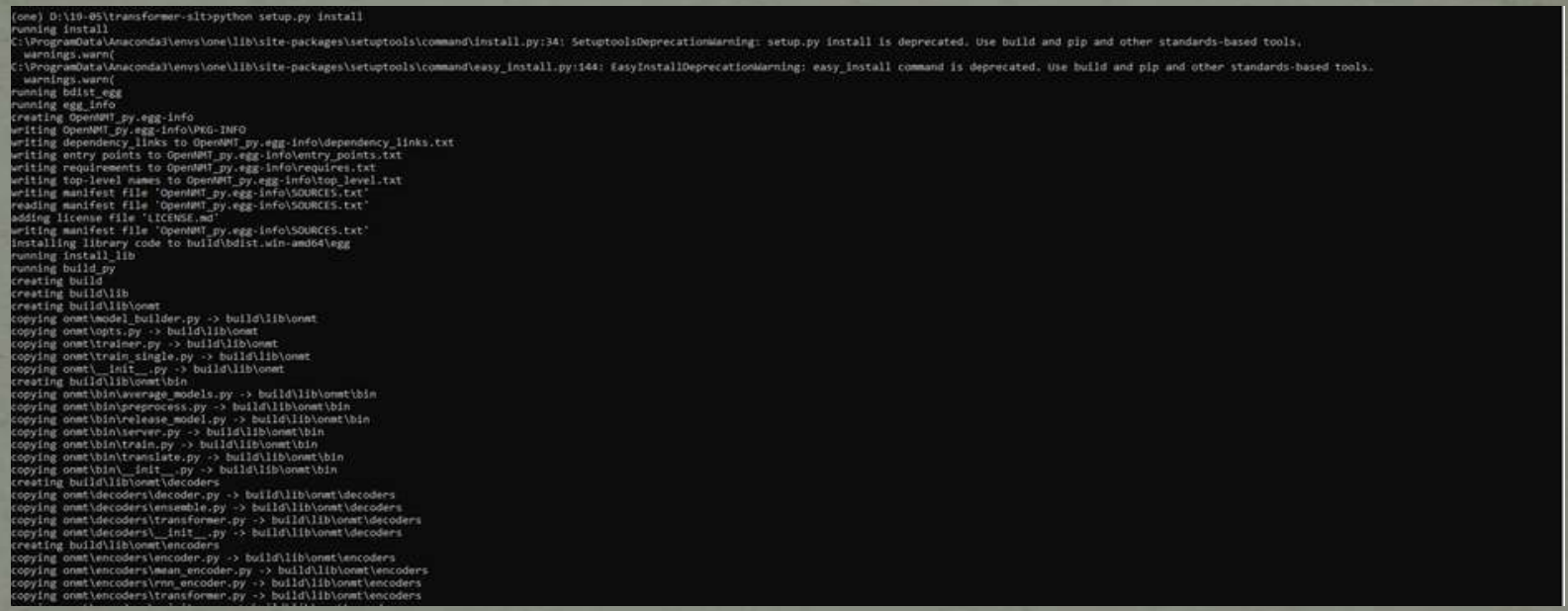
The inference time depends on the video length. In average, it takes around 8 seconds (25FPS) for a sign sentence. For a fair comparison, the inference time of 200 frames on a single GPU are evaluated. Compared with introducing an external VGG-11 based model for pose estimation, our self-contained branch saves around 44% inference time. It’s notable that our framework with the self-contained branch still shows slightly better performance than an off-the-shelf model. Role regularization is facilitated by the differentiable pose estimation branch and thereby the overfitting of neural networks is alleviated.

Scoring:

We are using BLEU.py to find the success ratio or the speed at which the translation has been done. This metric will show how far is the efficiency of one work from another. Since this is a Research based project, we use this scoring system to prove that this project has better efficiency and accuracy compared to the previous model.

Screenshots Of Working.

(1.) Setup.py installation:

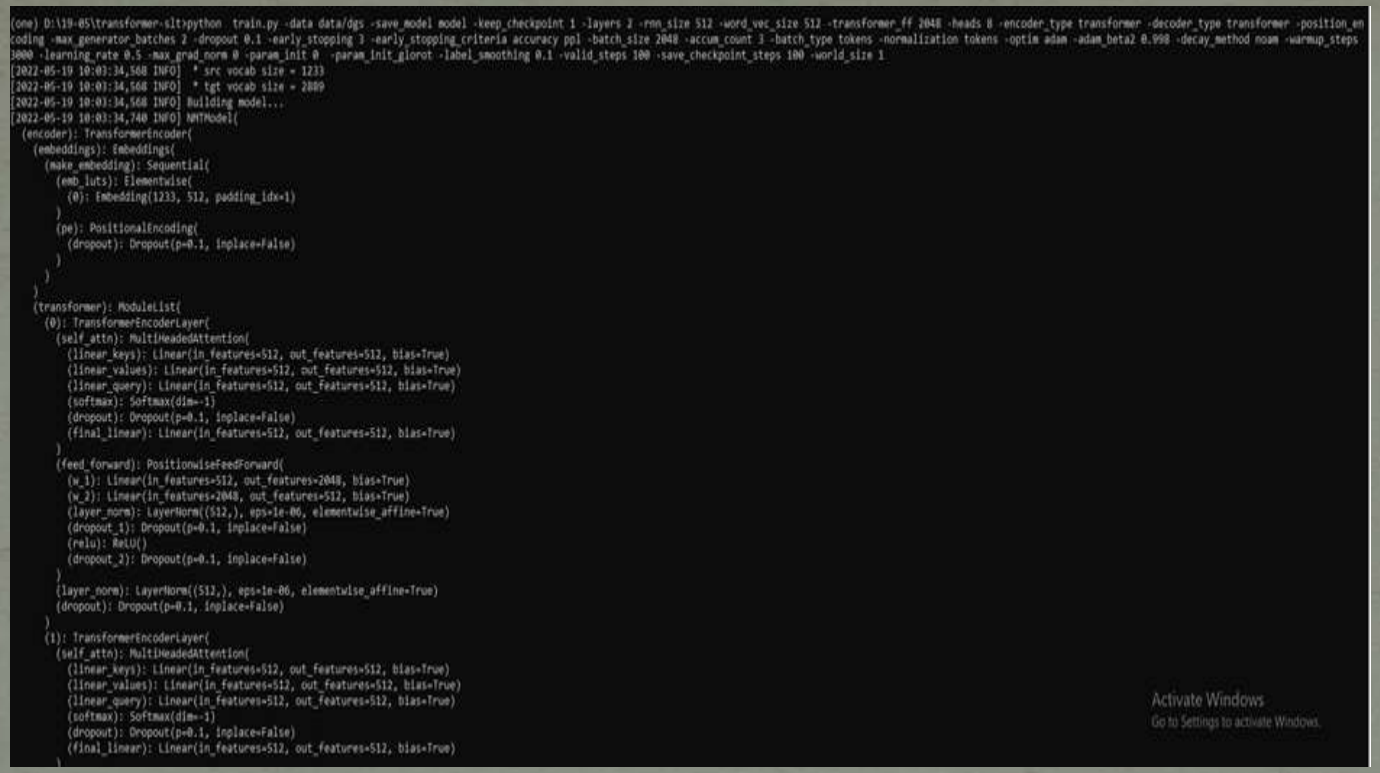


This is used to install necessary tools & packages for the application to process the Gloss text to Spoken language. The requirements will be installed in the virtual environment to perform the operation.

(2.) **Data Processing:**

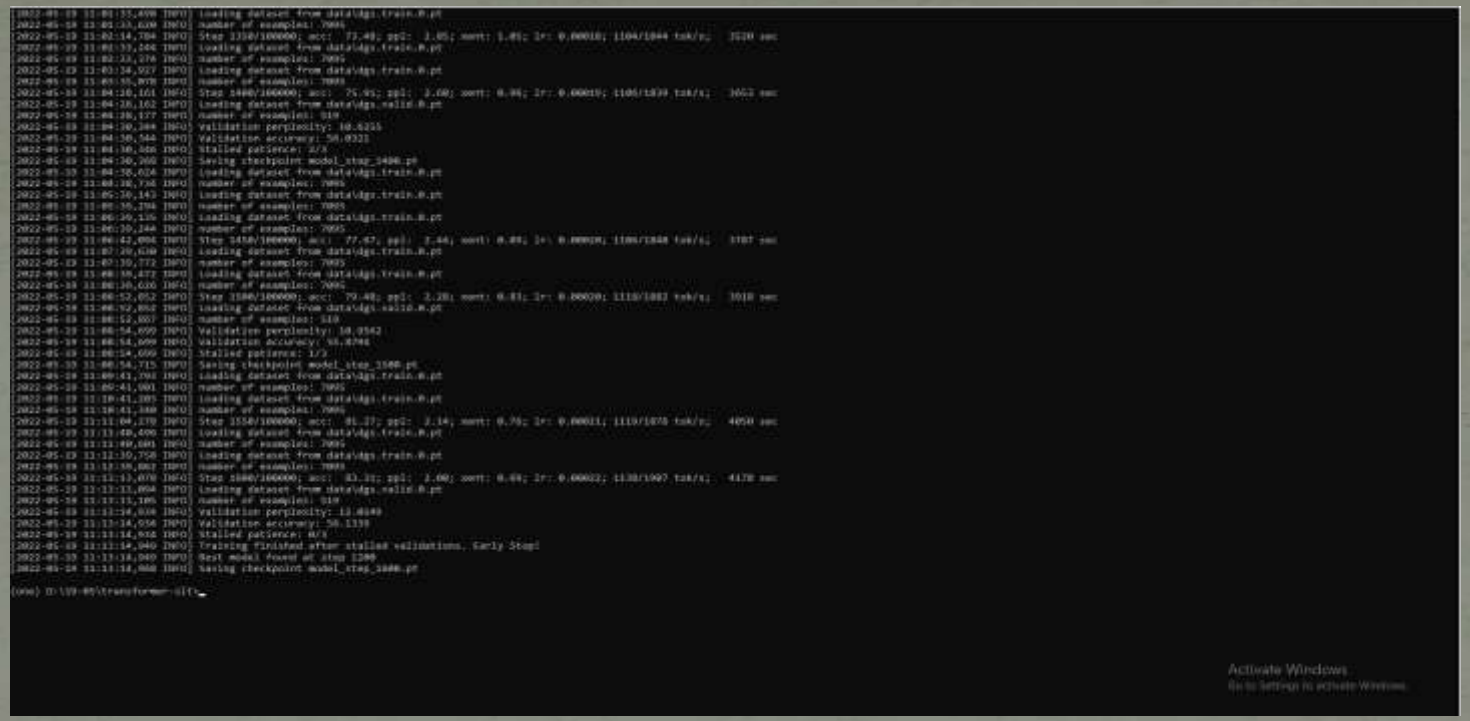
It performs the task of converting data from a given form to a much more desired form. In this,the function of Training and Validity is performed.

(3.) Training Of The Model:



This shows the models through which the words are trained for conversion of sign Language by using NLP(Natural Language Processing) & Hyper-parameter Tuning.

(4.) Output of the Trained model:



This shows that the training of the model has been completed and that the words have been learned.

Conclusion.

- Finally, We present a novel multi-cue framework for CSLR, which aims to learn spatial-temporal correlations of visual cues in an end-to-end fashion.
- In our framework, a spatial multi-cue module is designed with a self-contained pose estimation branch to decompose spatial multi-cue features.
- Moreover, we propose a temporal multi-cue module composed of the intra-cue and inter-cue paths, which aims to preserve the uniqueness of each cue and explore the synergy of different cues at the same time.
- A joint optimization strategy is proposed to accomplish multi-cue sequence learning.
- Extensive experiments on three large-scale CSLR datasets demonstrate the superiority of our STMC framework.

References:

- Kayo Yin; Graham Neubig. **Interpreting Language Models with Contrastive Explanations.**
- Amit Moryossef; Kayo Yin; Graham Neubig; Yoav Goldberg. **Data Augmentation For Sign Language Gloss Translation.**
- Kayo Yin; Jesse Read**. Better Sign Language Translation With STMC-Transformer.**
- Bahanau, D.; Cho, K.; and Bengio, Y. 2014**. Neural machine Translation by Jointly Learning to align and translate in ICLR.**
- Buehler, P,; Zisserman, A.; and Everingham, M. 2009. **Learning Sign Language by watching tv (using weakly aligned subtitles).** In CVPR.
- Chapelle, O., and Wu, M. 2010. **Gradient descent optimization of smoothed information retrieval metrics.** *Information Retrieval.*
- CihanCamgoz, N.;Hadfield, S,; Koller, o.; and Bowden, R. 2017. Subunets**: end-to-end hand shape and continuous sign language recognition**. In *ICCV.*
- CihanCamgoz, N.; Hadfield, s.; Koller, o.; Ney, H.; and Bowden, R. 2018. **Neural sign language translation**. In *CVPR.*
- Cooper, H., and Bowden, R. 2009. **Learning signs from subtitles: A weakly supervised approach to sign language recognition.** In *CVPR*
- Cui, R.; Liu, H.; and Zhang, c. 2017. **Recurrent convolutional neural networks for continuous sign language recognition byu staged optimization**. In *CVPR.*
- Cui, R.; Liu, H.; and Zhang, C. 2019. **A deep neural framework for continuous sign language recognition by iterative training**. *TMM* 21(7):1880-1891.
- Feichtenhofer, c.; Pinz, A.; and Zisserman, A. 2016. **Convolutional two-stream network fusion for video action recognition**.In *CVPR.*