سورة المجادلة - من آية ١١

Sign Language Identification Using Machine Learning Techniques

التعرف على لغات الاشاره باستخدام اساليب تعلم الاله

By

Eng. Ahmed Mahmoud Sultan

Under supervision

Prof Dr.Abdelmgeid Amin Ali
Department of Computer Science, Dean of Faculty of Computers and Information, Minia University

Prof Dr. Mohamed Sayed Kayed
Department of Computer Science, Dean of Faculty of Computers and Artificial Intelligence , Beni-Suef University

Dr. Walied Makram Mohamed
Department of Information Syetem, Faculty of Computers and Information, Minia University

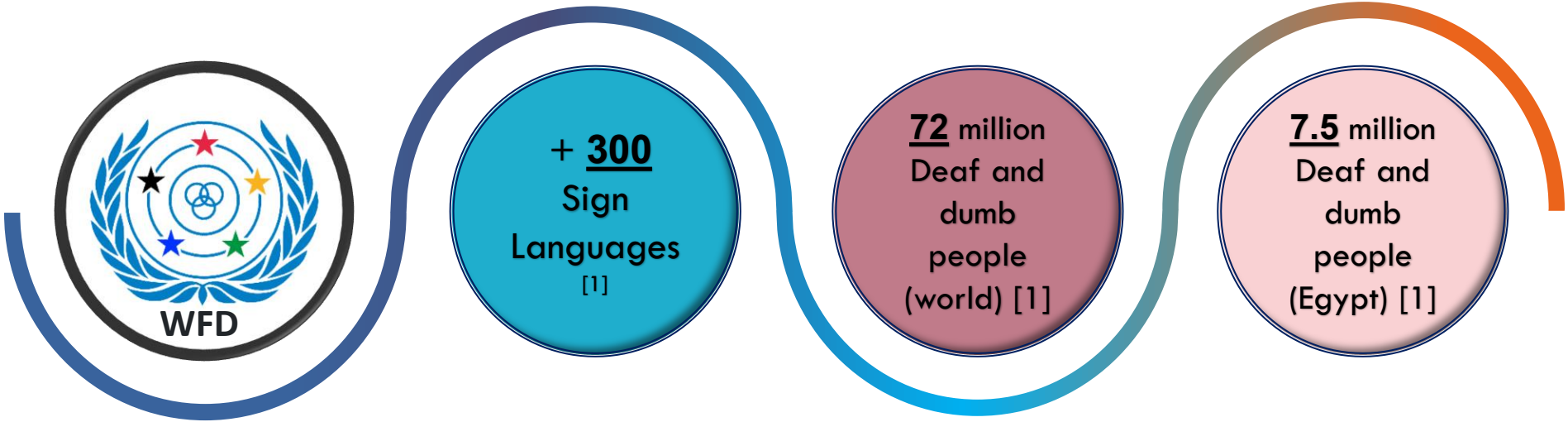
Agenda

- Overview.
- Problem Definition.
- Dataset.
- Model.
- Results.
- Conclusion and Future Work.
- Bibliography.

Research Papers

- "**Sign language identification and recognition: A comparative study**" Open Computer Science, pp. 191-210, 2022.
- "**Multiple Sign Language Identification using Deep Learning Techniques**" under review.

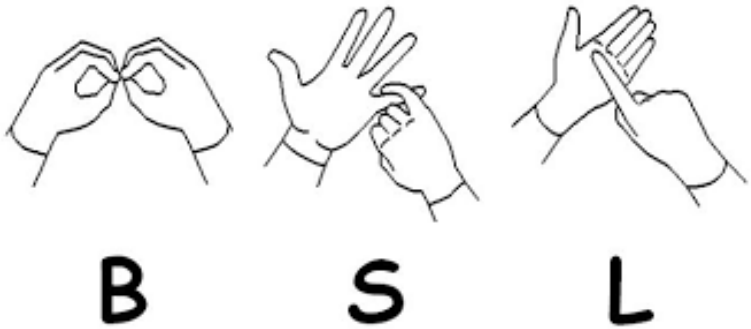
Overview


WFD
+ 300
Sign
Languages
[1]
72 million
Deaf and
dumb
people
(world) [1]
7.5 million
Deaf and
dumb
people
(Egypt) [1]

Overview

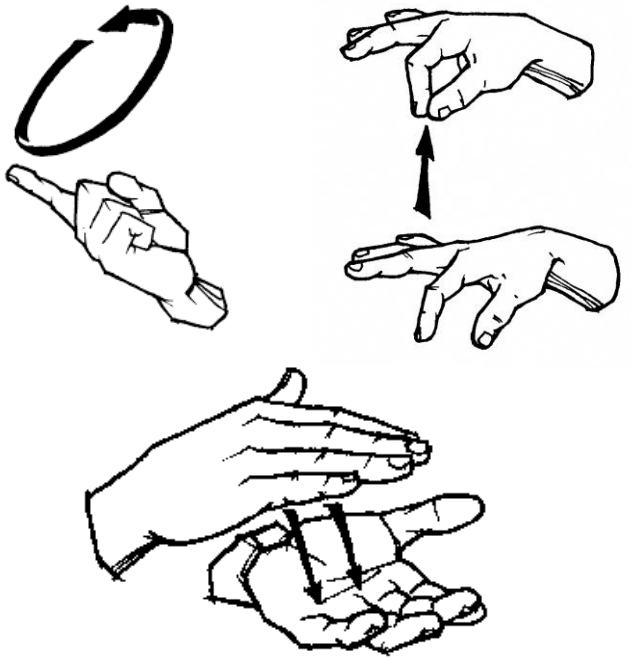
- There are many sign language across all the world. Different of these languages cannot be learned easily. So, using Machine Learning techniques will help to overcome and help more people to communicate with handicapper people.
- Classification of sign language such as Arabic Sign Language(**ArSL**) , American Sign Language (**ASL**), British Sign Language(**BSL**).
- Identify signs between three languages which may have the same shape.

Overview

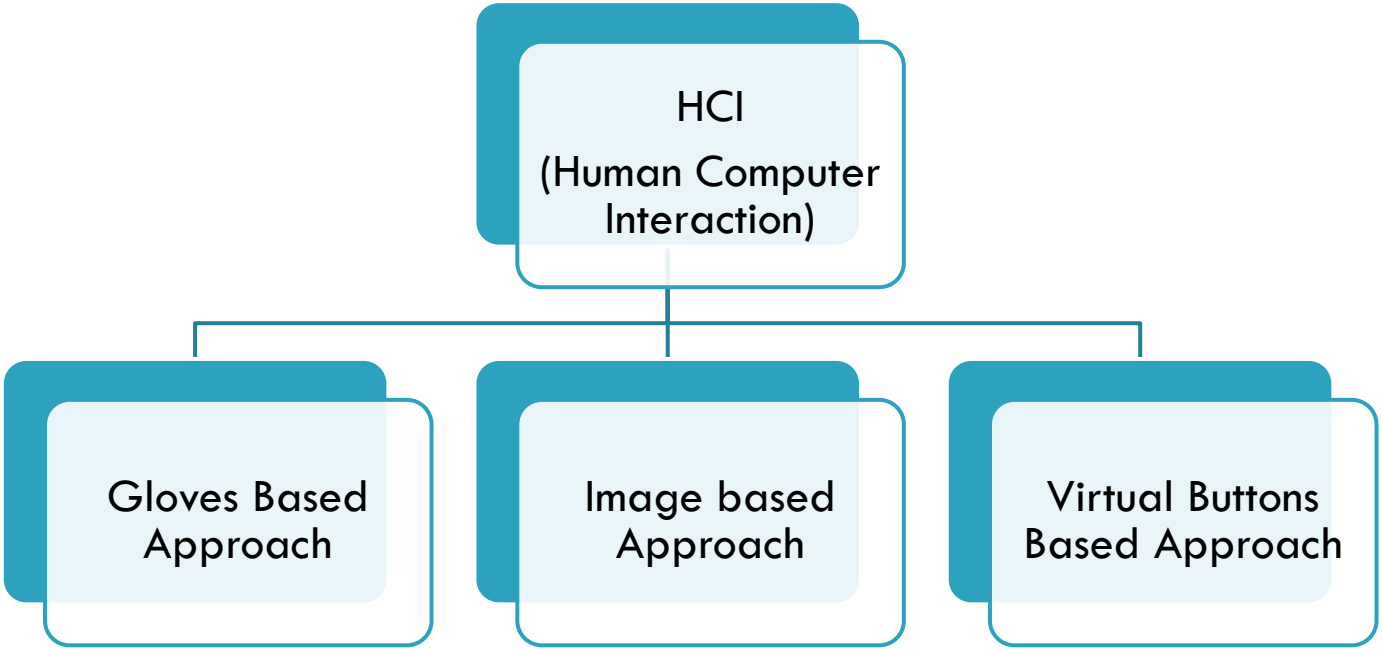


Overview

- **Hand gestures** enabling deaf people to communication during their daily lives rather than by speaking.
- A sign language is a language which, instead of using sound, uses visually transmitted gesture signs which simultaneously combine **hand shapes**, **orientation and movement of the hands**, **arms**, **lip-patterns**, **body movements and facial expressions** to express the speaker's thoughts.



Overview


HCI
(Human Computer Interaction)
Gloves Based Approach
Image based Approach
Virtual Buttons Based Approach

Overview

Gloves Based Approach

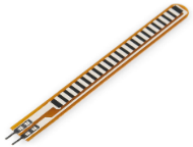
- This method employs **sensors** (mechanical or optical) attached to a glove that transduces finger flexions into electrical signals for determining the hand posture.
- Advantages and Disadvantages.

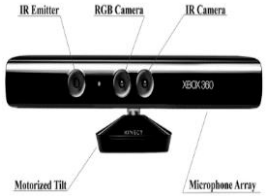
Example

Overview

Gloves Based Approach

- **Tilt Sensor:** used to measure slope and tilt with a limited range of motion.
- **Accelerometer Sensor:** -used to measure the rate of change of velocity.
- **Flex Sensor:** - it is a very thin and lightweight electric device, used for the measurement of bending or deflection.
- **Motion (proximity) Sensor:** - electrical device utilizes a sensor to capture motion.



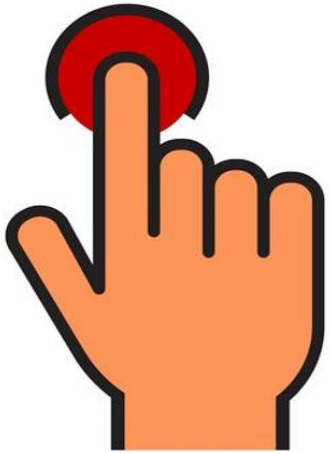


Overview

Virtual Button Based Approach

- **Virtual Button approach** : depends on a **virtual button** generated by the system and receives hand's motion and gesture by holding and discharging individually.

Example

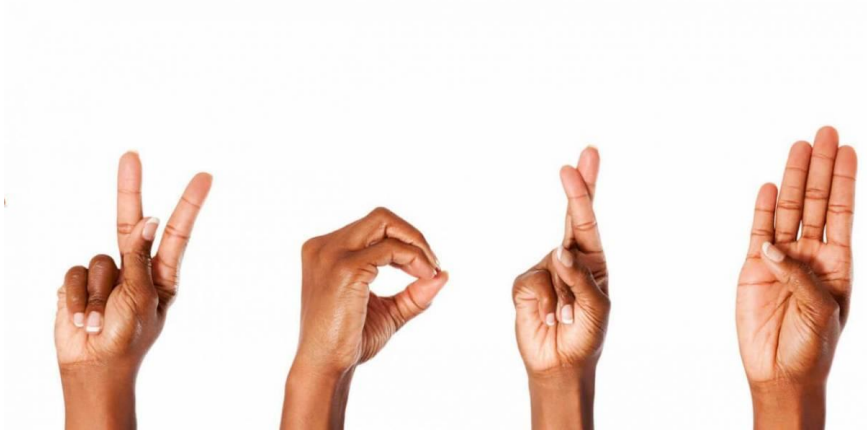


Overview

Vision Based Approach

- **Vision-based approach :** this approach required only a **Camera** to capture a person's movements with a clear background without any gadgets. Previous gloves required an accompanying camera to register the gesture but does not work well in lightning conditions.

Example



Agenda

- Overview.
- **Problem Definition.**
- Dataset.
- Model.
- Results.
- Conclusion and Future Work.
- Bibliography.

Problem Definition.

- The Need for a well-structured benchmark **dataset** of different sign language such as (**ArSL** – **ASL** – **BSL**) (which has variance in background- illumination- different hand signers).
- The main purpose of the study is to identify sign language letters and define the language type that this letter belongs to. There are many **conferences** That signers need to communicate with handicapped people, we need a system that can **Identify** more than one time language. We used deep learning techniques such as **LeNet**, **VGG-16**, and **CapsNet**. These algorithms give Promising results.

Agenda

- Overview.
- Problem Definition.
- **Dataset**.
- Model.
- Results.
- Conclusion and Future Work.
- Bibliography.

Dataset

- Dataset collected from participants who have more **awareness** of SLs.
- Dataset was collected from different **environments,** for more variance.
- Dataset has only **Vision-based** approach.
- Dataset consist of images which depict the **Alphabets.**

DataSet Cont… (ASL)[12]

Size	No of participants	Devices	Content (Video/image/Gloves)
201 (Video)	Not mentioned	Digital Camera [Black – white]	Video
5829 (Phrase)	30	Gloves with accelerometer	Gloves
25,000 videos	100	-	Video
21,083 videos	119	-	Video

DataSet Cont… (ArSL)[12]

Size	No of participants	Devices	Content (Video/image/Gloves)
40 (Phrase)	Not mentioned	DG5-VHand data gloves & Polhemus G4	Gloves
20 (Word)	Not mentioned	Digital-Camera	Videos
75,300 videos	3	Multi-modal Microsoft Kinect V2	Video

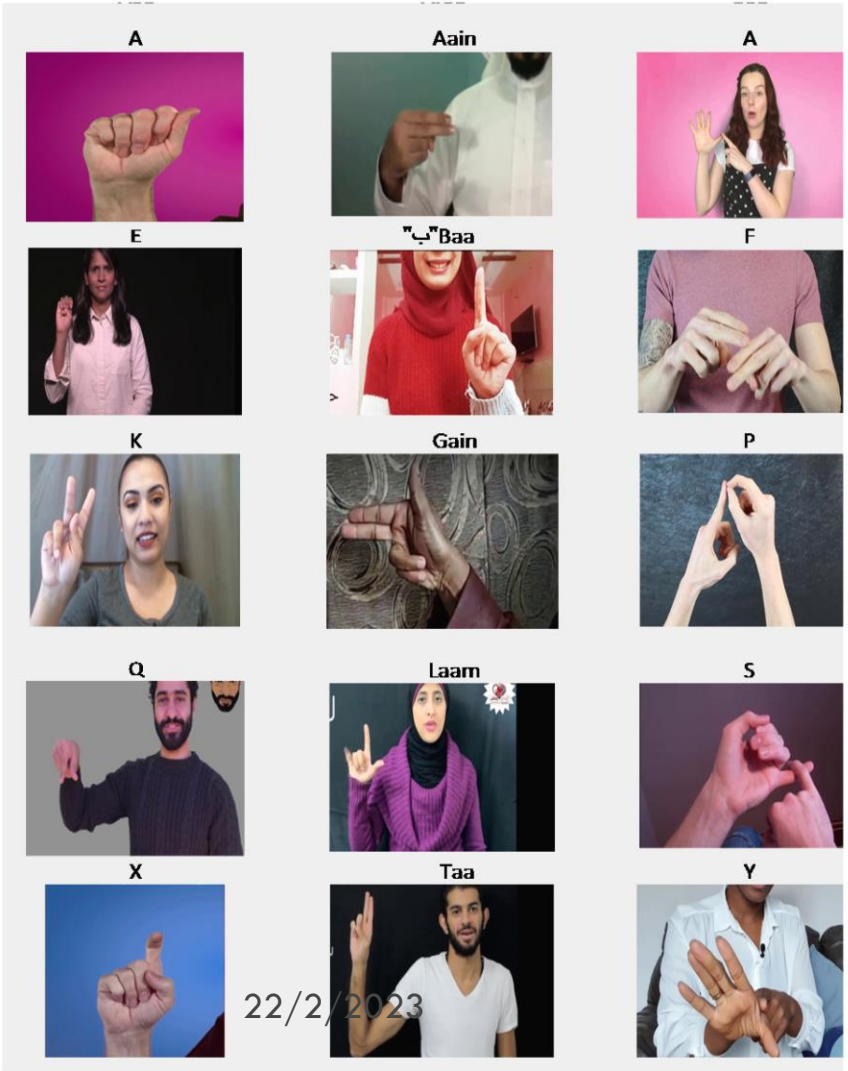
DataSet Cont… (BSL & GSL)[12]

Size	No of participants	Devices	Content (Video/image/Gloves)
1400 (image)	20	Web Cam	1400 (image) Image
6.44 Hours	7 (Native Greek Signers)	Intel RealSense D435 RGB+D camera	Video
~16 hours	10 Greek 9British	Digital Camera	Video

Our Constructed Dataset

SL	Hands	Size(image)	No of signs	No of Signers
ArSL	Single hand	41,959	29	16
ASL	Single hand	38,483	26	14
BSL	Double hand	61,120	26	26

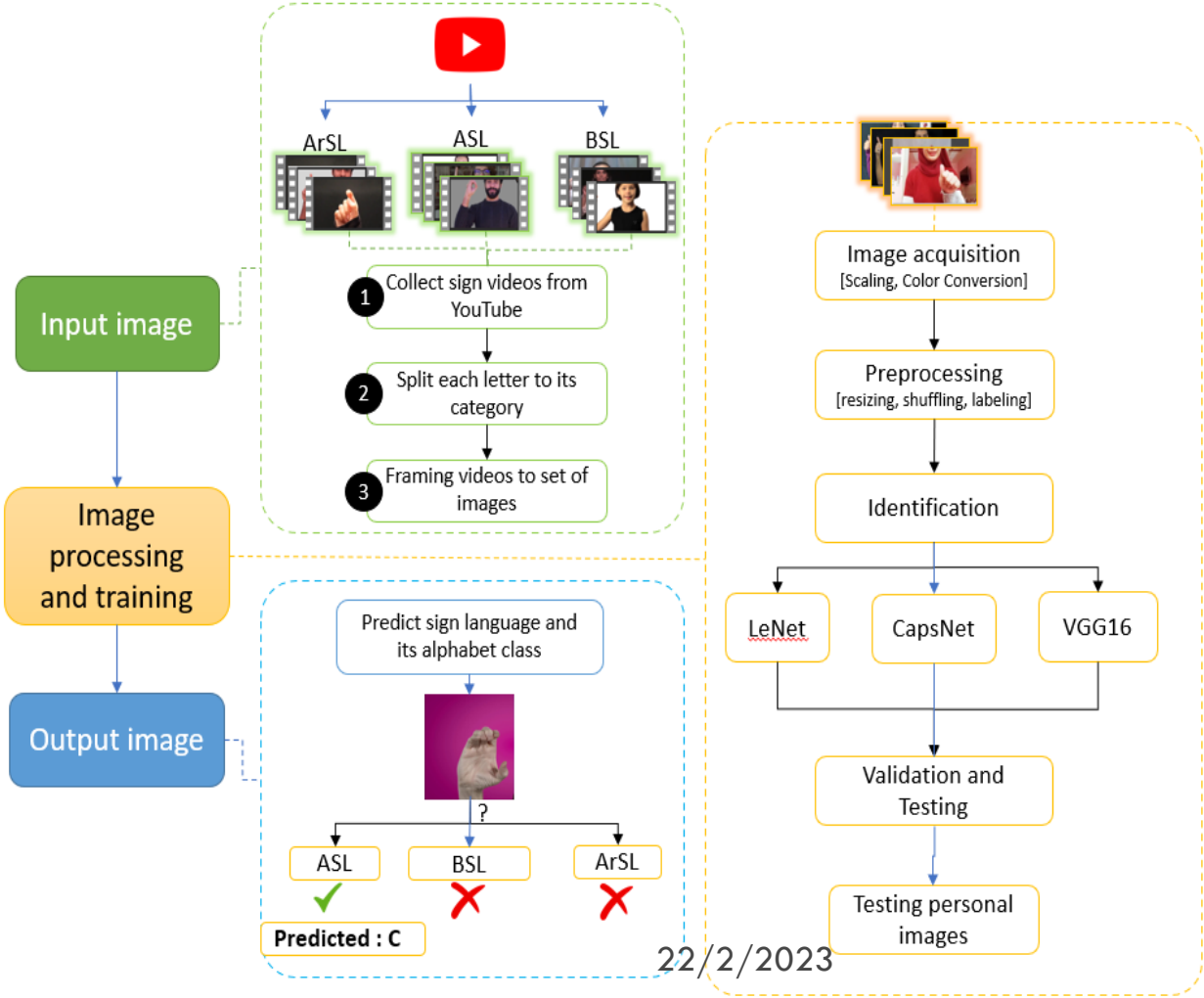
Dataset Samples



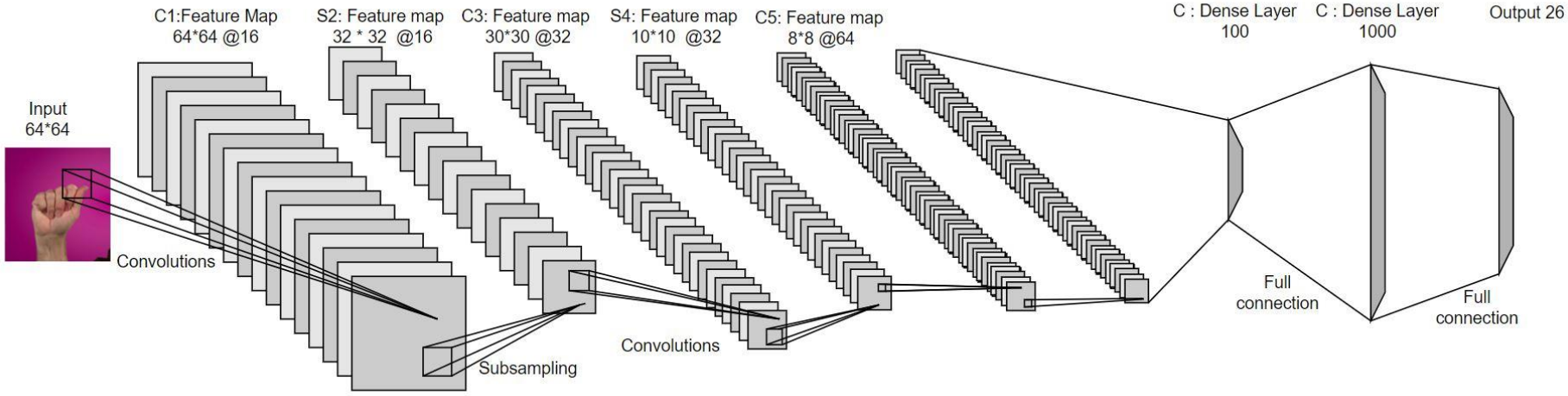
Agenda

- Overview.
- Problem Definition.
- Dataset.
- **Model.**
- Results.
- Conclusion and Future Work.
- Bibliography.

Model


Input image
Image processing and training
Output image
ArSL
ASL
BSL
1 Collect sign videos from YouTube
2 Split each letter to its category
3 Framing videos to set of images
Predict sign language and its alphabet class
?
ASL
BSL
ArSL
Predicted : C
Image acquisition
[Scaling, Color Conversion]
Preprocessing
[resizing, shuffling, labeling]
Identification
LeNet
CapsNet
VGG16
Validation and Testing
Testing personal images

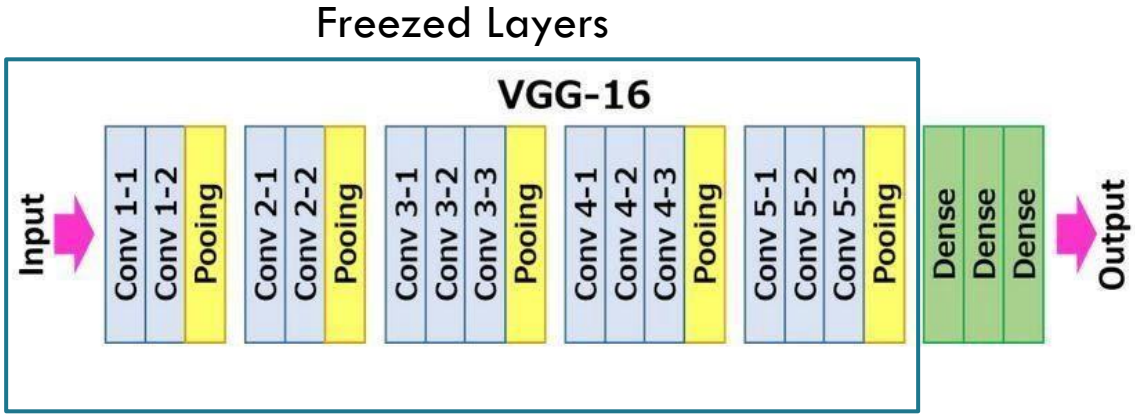
LeNet Model


C1:Feature Map
64*64 @16
S2: Feature map
32 * 32 @16
C3: Feature map
30*30 @32
S4: Feature map
10*10 @32
C5: Feature map
8*8 @64
C : Dense Layer
100
C : Dense Layer
1000
Output 26
Input
64*64
Convolutions
Subsampling
Convolutions
Full
connection
Full
connection

VGG-16 Model

- VGG is based on classical CNN architecture. VGG stands for Visual Geometry Group.Vgg16 [2] is a pretrained model.
- VGG-16 could extract more features than AlexNet.
- VGG16 was applied to ImageNet dataset [3] which has 1000 class of different categories, reaching a test accuracy of 92.7%.

VGG-16 Model


Freezed Layers
VGG-16
Input
Conv 1-1
Conv 1-2
Pooing
Conv 2-1
Conv 2-2
Pooing
Conv 3-1
Conv 3-2
Conv 3-3
Pooing
Conv 4-1
Conv 4-2
Conv 4-3
Pooing
Conv 5-1
Conv 5-2
Conv 5-3
Pooing
Dense
Dense
Dense
Output

Drawbacks of CNN architectures

1. It tends to **lose** a lot of information(MaxPool, MinPool, AvgPool).
2. Pooling layer **loses** the required **spatial information** about the **rotation**, **location**, **scale**.
3. Another drawback of the pooling layer is that if the **position** of the object is slightly changed the activation doesn't seem to **change** with its proportion, which leads to **good accuracy** in terms of image classification but **poor in performance**, if you want to **locate** exactly where the **object** is in the **image**.

CapsNet Model[4]

- A ***capsule*** is a collection or group of **neurons** that stores different information about the object it is trying to identify in a given image.
- Information mostly about its position, rotation, scale, and so on in a high dimensional vector space (8 dimension or 16 dimension).
- **Rendering** in graphics and its relationship with **'inverse graphics'** by humans.

CapsNet Model[4]

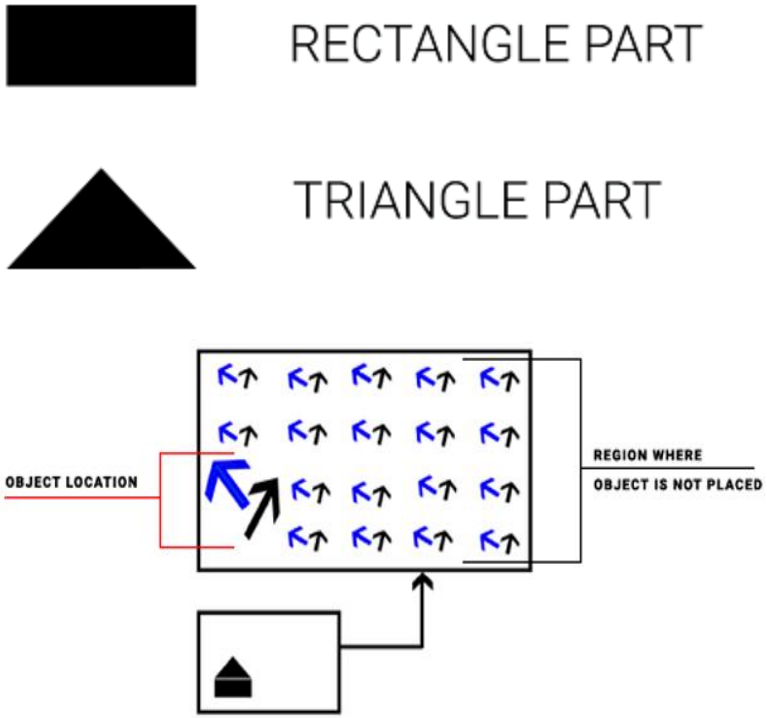
- Primary capsules
  - Convolution
  - Reshape
  - Squash
- Higher layer capsules
  - Routing by agreement
- Loss calculation
  - Margin loss
  - Reconstruction loss

CapsNet model

- Primary Capsules.
  - The idea behind caps net model.
  - What did you see?
    - For the black objects ! Which is boat and which is house?
    - A sign of hand(ASL : B)! Could you recognize it as a boat or a cat?

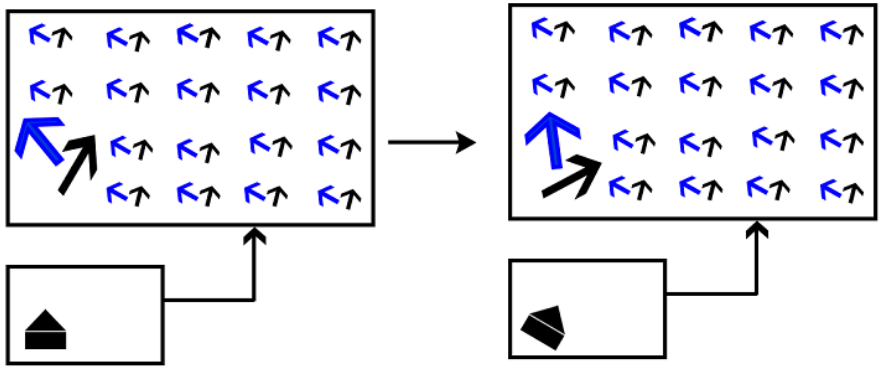
CapsNet model

- Capsules representing the triangle and rectangle will be constructed.
- The output of these capsules is represented with the help of arrows.
- The black arrows representing the rectangle's output.
- The blue arrows representing the triangles.



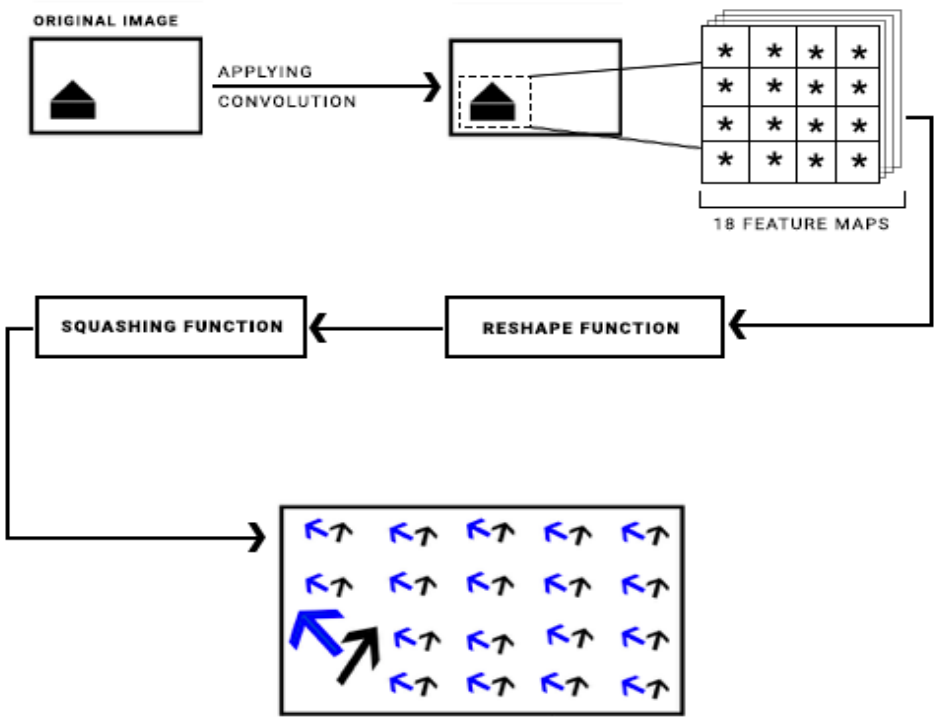
CapsNet model (equivariance)

- if we slightly **rotate** the object in our input image, the arrows representing these objects will also slightly **rotate** with proportion to its input counterpart. This is known as **'equivariance'.**
- This enables the capsule networks to **locate** the object in a given image with its precise location, scale, rotation, and other positional attributes associated with it.



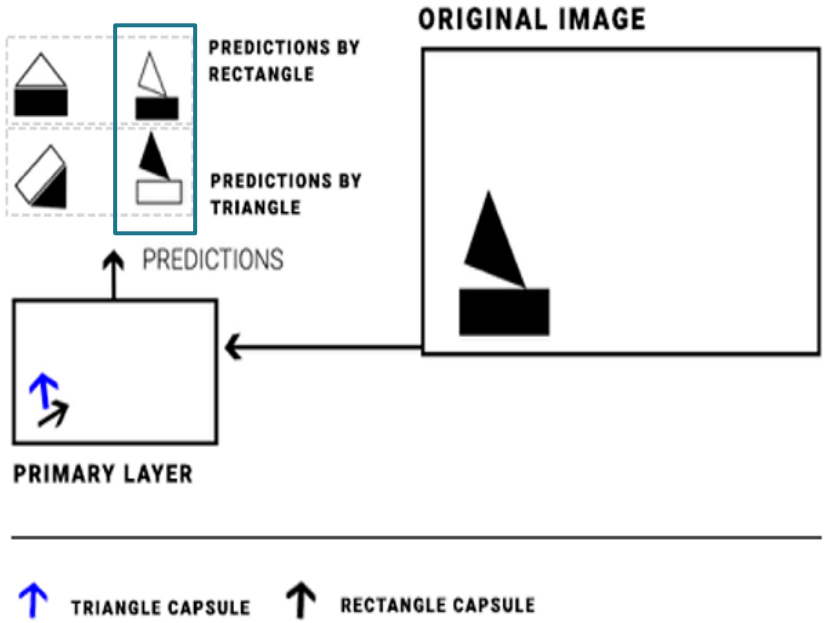
CapsNet model

- This is achieved using three distinct processes:
  - Convolution
  - Reshape function
  - Squash function
    - Should be between 0,1 ,as it is the probability of existence.



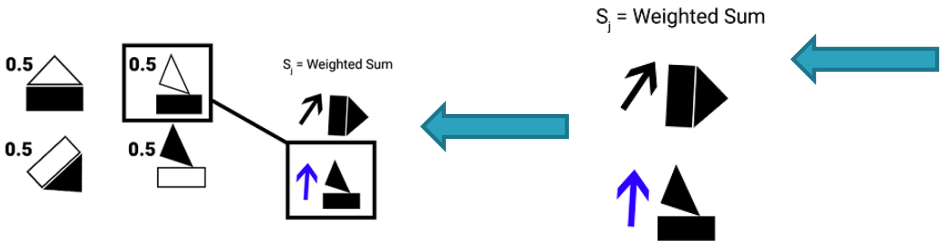
CapsNet model (Routing By Agreement)

- **‘Routing By Agreement’.**
- **Benefit:-** Once the primary capsules agree to select a certain higher-level capsule, there is no need to send a signal to another capsule in another higher layer.



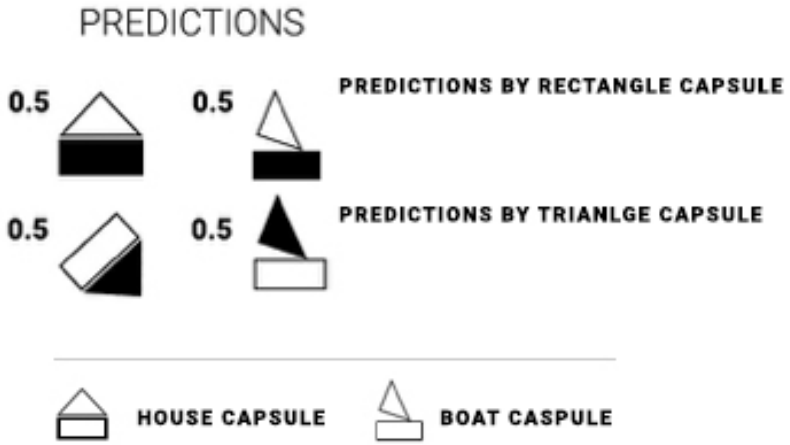
CapsNet model (Higher Layer)

- The first step the higher capsule layer takes to calculate its own output is to set up something called '**routing weights**'.
- after assigning the Softmax output to the predictions, it calculates the weighted sum to each capsule in this higher layer. (Round one)



$B_{ij} = 0$ for all i and j

softmax(B_i)



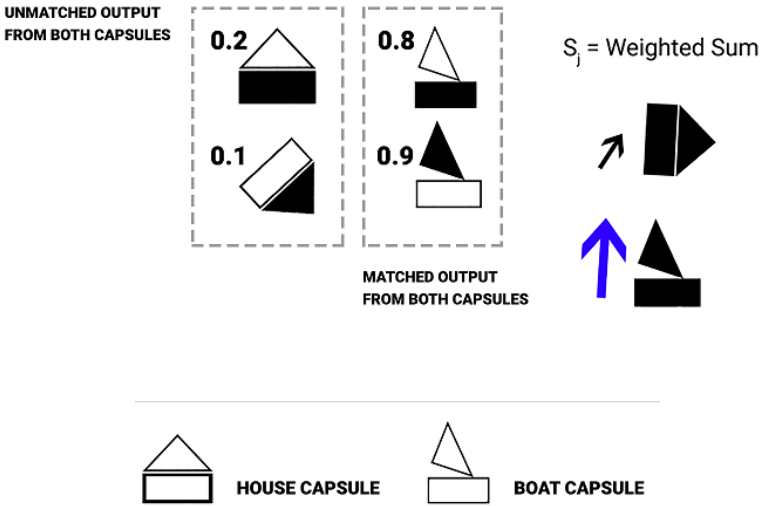
CapsNet model (Higher Layer)

- we again calculate another routing weight for the **next round** by scalar product of the prediction and the actual output of the layer and adding it to the existing routing weight.

$$B_{ij} \mathrel{+}= \hat{U}_{ij} + V_j$$

$\hat{U}_{ij}$ (Prediction by primary layer)

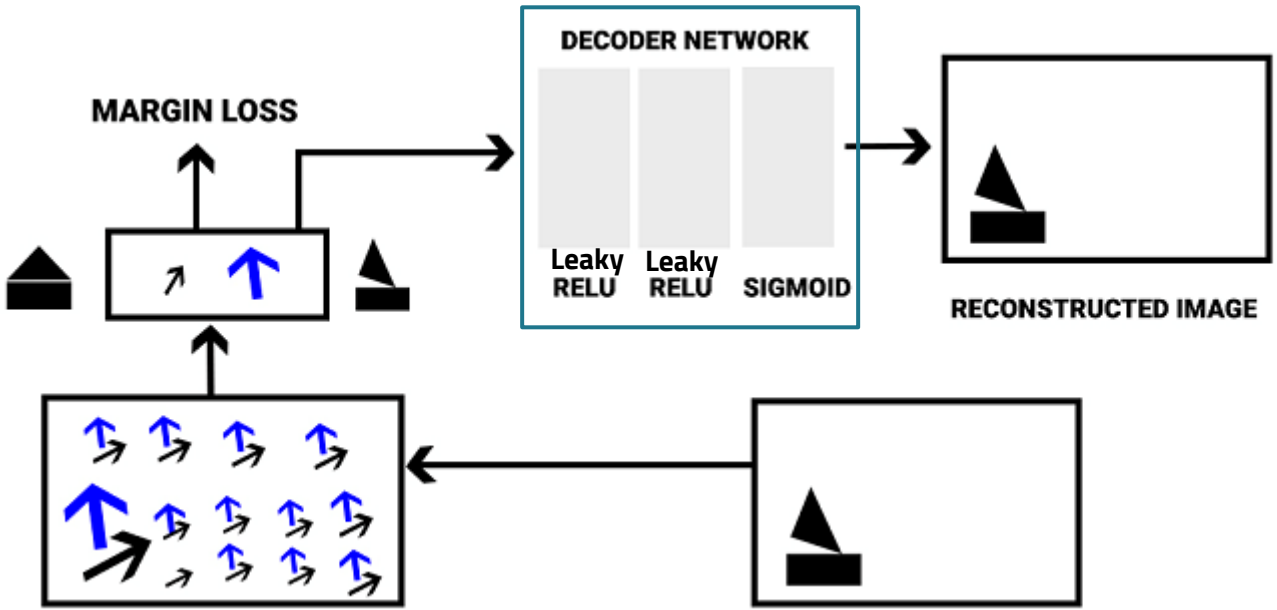
V_j (Actual output by the higher layer)



CapsNet model

- Loss Function
  - Now that we have made a decision on what object is in the image using the 'routing by agreement' method, you can perform classification.
  - margin loss was used to calculate the class probability of multiple classes to create such an image classifier.
  - $\mathbf{V_k}$ is the length of the output vector of that class **K** object. Now, if that object of class **K** is present then its squared value should not be less than 0.9;
    - $|\mathbf{V_k}|^2 \geq 0.9$ Object exist.
    - $|\mathbf{V_k}|^2 \leq 0.1$ Object not exist.

CapsNet model

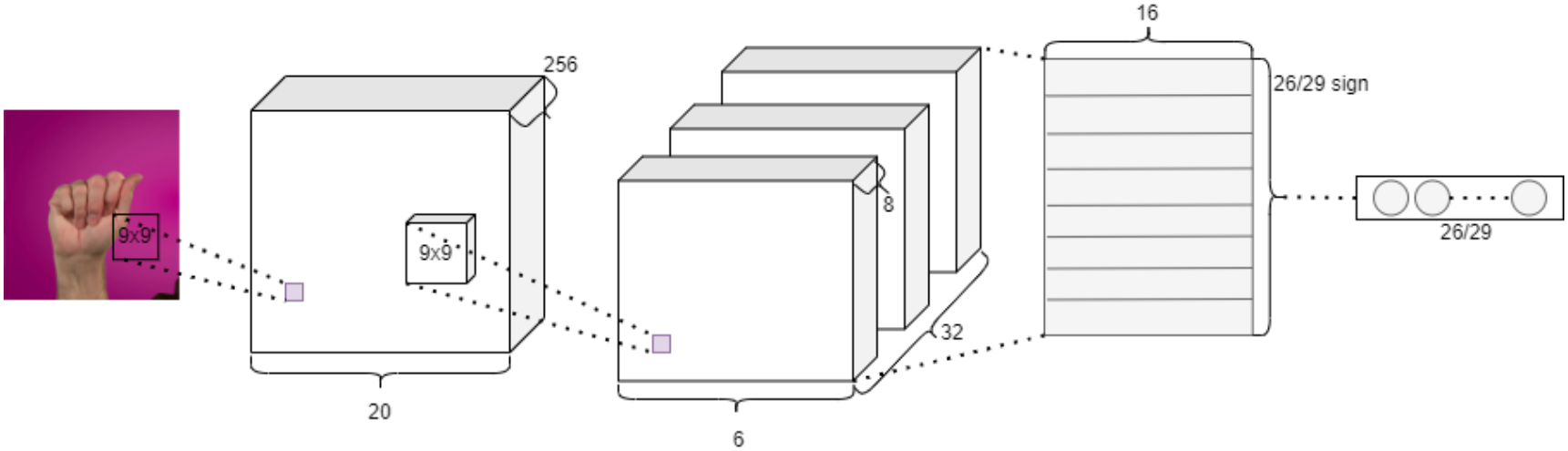


Reconstruction Loss = (Reconstructed Image – Input Image)^2

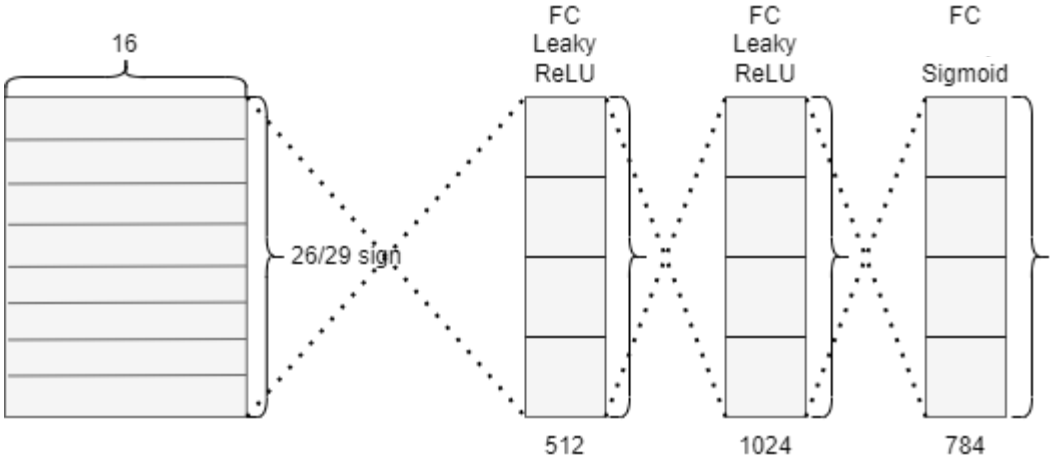
Total Loss = Margin Loss + alpha * Reconstruction Loss

Dimension perturbations

CapsNet Architecture



CapsNet Architecture



Agenda

- Overview.
- Problem Definition.
- Dataset.
- Model.
- **Results.**
- Conclusion and Future Work.
- Bibliography.

Results(Hyperparameters)

Hyperparameters	LeNet Model	VGG16	CapsNet
Activation Function	ReLU/ SoftMax	ReLU/ SoftMax	Leaky ReLU
Learning Rate	-	$1e^{-5}$	-
Epochs	150	25	10
Batch Size	150	128	50
Loss Function	Sparse categorical cross entropy	Categorical Cross Entropy	Margin loss (MSE)
Optimizer	Adam	RMSprop	Adam

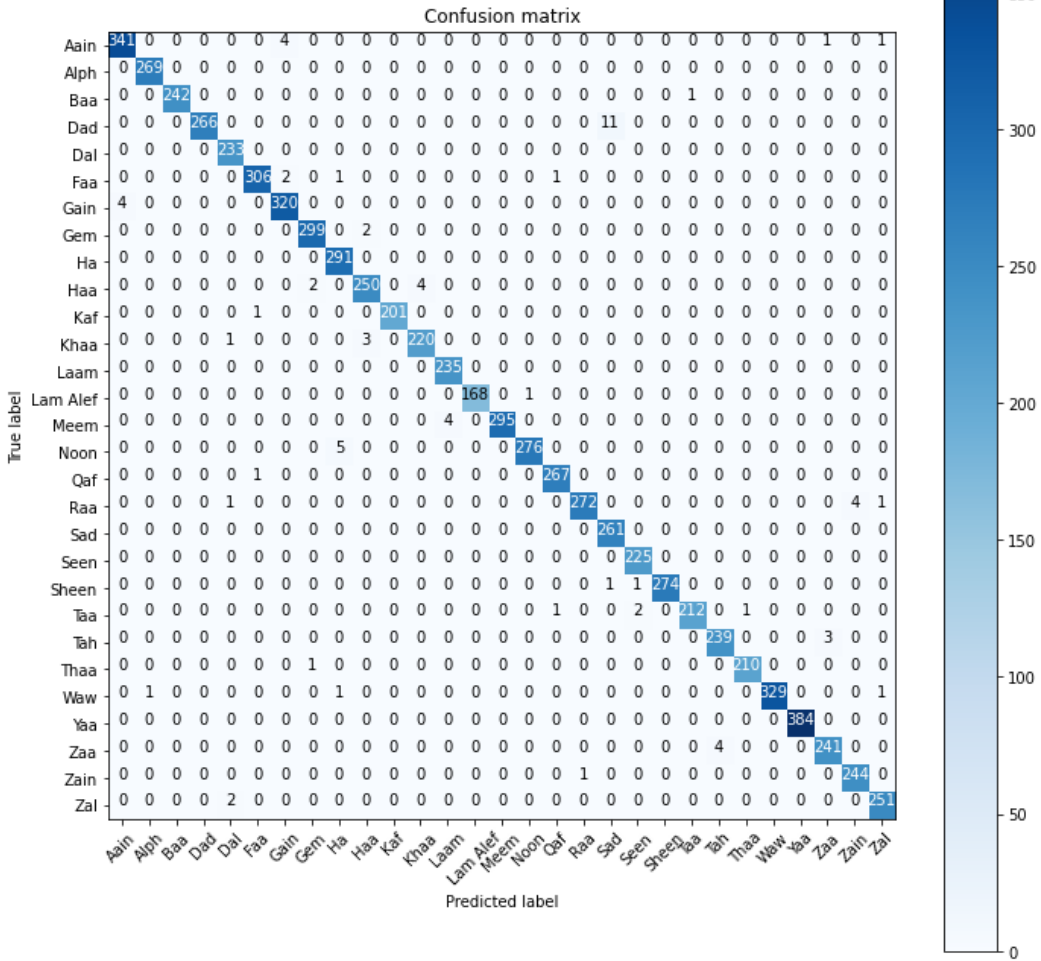
Results

Author	Method	Dataset	Device	Size (image)	Classes	batch size /epochs	Loss	Training Accuracy
Our own	LeNet	ArSL	Camera	26,100	29	50 / 50	0.1600	94.95%
		ASL	Camera	13,000	26	150 / 150	0.0468	98.43%
			Camera	13,000	26	100 / 100	0.0788	97.45%
		BSL	Camera	13,000	26	150 / 150	0.0387	**98.79%**
			Camera	18,200	26	150 / 150	0.1536	95.02%
			Camera	50,906	26	150 / 150	0.1124	96.54%
	VGG-16	ArSL	Camera	41,959	29	128 / 25	0.0358	99.05%
		ASL	Camera	38,483	26	128 / 25	0.0649	98.50%
		BSL	Camera	61,120	26	128 / 25	0.0135	**99.69%**
	CapsNet	ArSL	Camera	26,100	29	50 / 10	0.026618	98.4848%
		ASL	Camera	23,400	26	50 / 10	0.033218	98.4286%
		BSL	Camera	36,010	26	50 / 10	0.012218	**99.5652%**

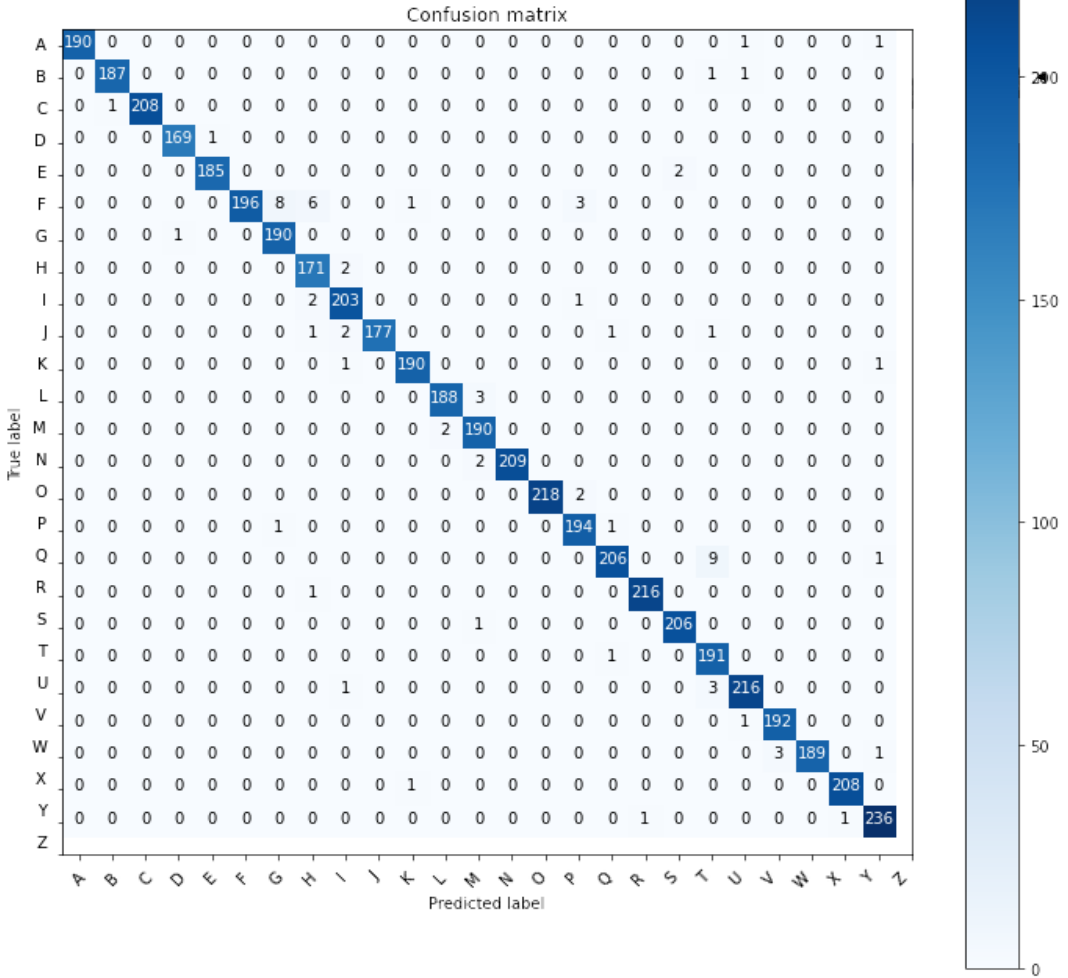
Comparable Results

Author	Method	Dataset	Device	Size (image)	Classes	batch size /epochs	Loss	Training Accuracy
El zaar [5]	VGGNET	ASL	Camera	58,114	29	32 / 40	-	99%
		ArASL	Camera	54,049	32	32 / 40	-	98%
		IrSL	Camera	58,120	26	32 / 40	-	99%
Nguyen [6]	VGG-16	ASL	Camera	5,391	26	- / 50	-	**99.902%**
Alsaadi [7]	VGG-16	ArSL	Camera	54,049	32	32 / 30	0.1957	94.05%
Abu-Jamie [8]	VGG-16	ASL	Camera	43,500	29	- / 20	0.0020	**99.99%**
Bilgin [9]	LeNet	ASL	Camera	34,627	24	128 / 30	-	82%
	CapsNet		Camera			128 / -		95%
Xiao [10]	CapsNet	ASL	Camera	34,627	34	50 / -	-	-
Suri [11]	CapsNet	InSL	IMU Device	2000 sentence	20 sentences	- / -	0.01	**99.72%**

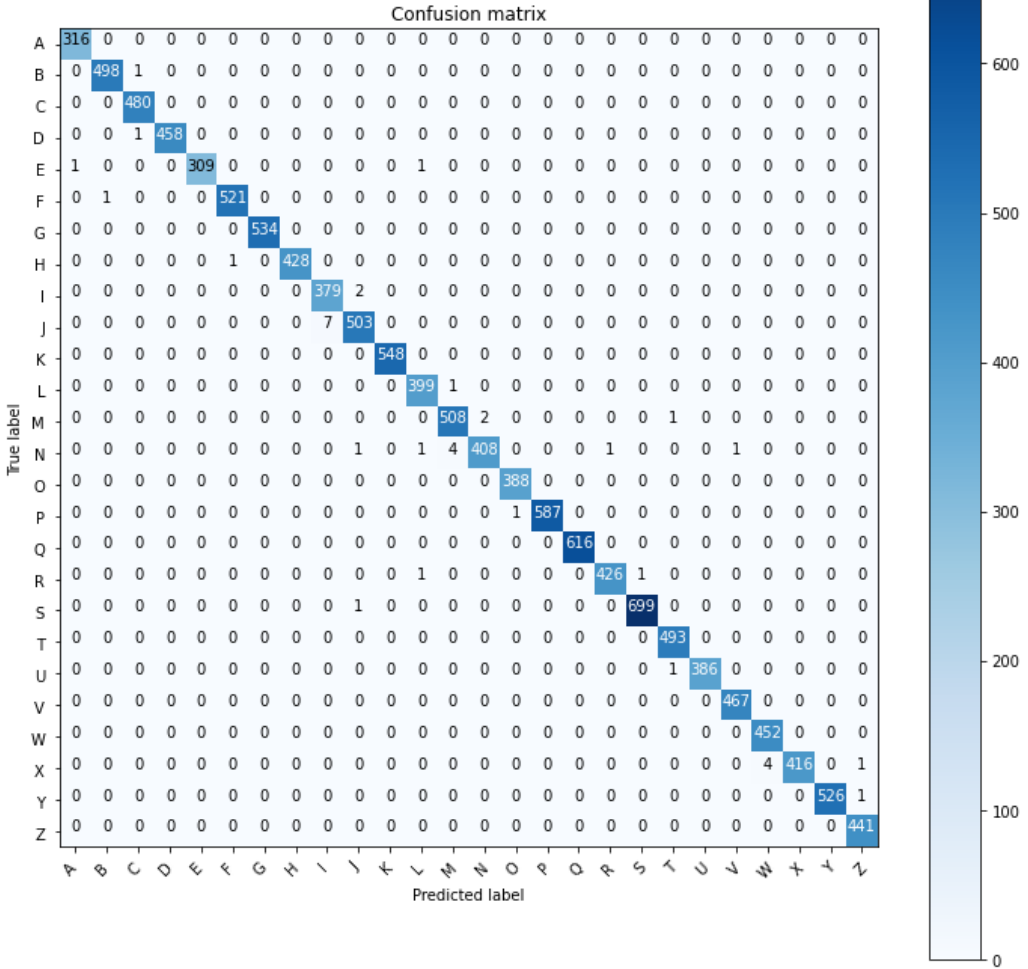
Confusion Matrix of ArSL (VGG-16)



Confusion Matrix of ASL (VGG-16)



Confusion Matrix of BSL(VGG-16)



Agenda

- Overview.
- Problem Definition.
- Dataset.
- Model.
- Results
- **Conclusion and Future Work.**
- Bibliography.

Conclusion

- Different approaches of sign languages led to different models to help identify and recognize hand gestures.
- Vision based approach is the best one, because signer is free of any portable devices.
- Deep learning algorithms results in high accuracies rather than traditional machine learning algorithms to recognize and identify signs.
- A new benchmark was built and get considerable results.
- LeNet, VGG-16, and CapsNet were used to identify each sign, and identify each language.
- We got the best accuracy using VGG-16.
- We could get best results using CapsNet, but we need more GPU's capacity.

Future Work

- As a future work, U-net model can be used to pre-process incoming images to be classified more efficiently.
- Mask R-CNN is a state-of-the-art framework for Image Segmentation tasks.
- Recognize and classify real time sign language images and videos.
- Using Avatar to visualize signs during interaction between deaf people and ordinary people.

Agenda

- Overview.
- Problem Definition.
- Work Plan.
- Dataset.
- Image Captioning Model.
- Conclusion and Future Work.
- **Bibliography.**

Bibliography

- [1] https://www.un.org/ar/observances/sign-languages-day.
- [2] S. Karen and Z. Andrew, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.
- [3] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009.
- [4] S. Sabour, N. Frosst and H. Geoffrey E, "Dynamic routing between capsules," Advances in neural information processing systems, 2017.
- [5]A. El Zaar, N. Benaya and A. El Allati, "Sign language recognition: High performance deep learning approach applyied to multiple sign languages," E3S Web of Conferences, 2022.
- [6] H. T. Nguyen, L. T. Pham, T. T. Mai, T. K. Vo and T. T. Dien, "Letter recognition in hand sign language with VGG-16," Intelligent Systems and Networks, pp. 410-417, 2022.

Bibliography

- [7] M. M. Kamruzzaman, "Arabic sign language recognition and generating Arabic speech using convolutional neural network," Wireless Communications and Mobile Computing, vol. 2020, pp. 1-9, 2020.
- [8] M. Varsha and C. S. Nair, "Indian sign language gesture recognition using deep convolutional neural network," 2021 8th International Conference on Smart Computing and Communications (ICSCC), 2021.
- [9]M. Bilgin and K. Mutludogan, "American sign language character recognition with capsule networks," 2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), 2019.
- [10] H. Xiao, Y. Yang, K. Yu, J. Tian, X. Cai, U. Muhammad and J. Chen, "Sign language digits and alphabets recognition by Capsule Networks," Journal of Ambient Intelligence and Humanized Computing, vol. 13, no. 4, pp. 2131-2141, 2021.
- [11] K. Suri and R. Gupta, "Continuous sign language recognition from wearable Imus using deep capsule networks and game theory," Computers & Electrical Engineering, vol. 78, pp. 493-503, 2019.
- [12] A. Sultan, W. Makram, M. Kayed and A. A. ALi, "Sign language identification and recognition: A comparative study," Open Computer Science, pp. 191-210, 2022.

Thanks and Acknowledgement