IMPLEMENTATION OF THE MOBILENET SSD ALGORITHM FOR SIGN LANGUAGE INTERPRETER SYSTEMS

Mualif Ulil Misbakh, Imam Husni Al Amin

Stikubank University, Semarang 50241, Indonesia

Abstract

In March 2020, WHO published an article entitled "Deafness and hearing loss", in which it stated that more than 5% of the world's population, or around 430 million people lost the ability to hear, and about 34 million of them were children. Sign language is a way or means for deaf and speech-impaired individuals to communicate and stay connected with the rest of the world and express ideas, from which there is a great need for efficient and cost-effective translation software or tools in the modern world to accurately understand what individuals with disabilities want to express, in this research project produced a software system to translate sign language into a natural language such as Indonesian for real-time communication, The translated data will imply the alphabet related to movements captured directly through the camera by utilizing SSD-MobileNet algorithm that has a light computing load but with high accuracy results up to more than 98%.

Keywords: SSD-MobileNet, Object detection, Sing language

1. Introduction

This time we live in an ever-changing society, which causes the lives of individuals like us to become younger and happier. It can be said that it is the absorption of knowledge from those who have first formed new ideas and shared them that makes each person play an important role in developing a better social life. But nowadays there are also many people out there who have a lot of difficulty in acquiring or expressing new ideas and communicating with others, especially for people who have disabilities in listening or speaking who struggle to overcome challenges and difficulties every day to be able to communicate with the people around them.

Until now, there is no data or statistics that show the exact number of people with deaf and speech disabilities. But in March 2020 the WHO has released an article entitled "Deafness and hearing loss" which estimates that there are about 5% of the world's population, or around 430 million people need rehabilitation for their hearing loss, and among them, there are about 34 million of them children, and it is also estimated that by 2050 there will be 900 million people with hearing loss.

Research conducted by Gumelar G, et al (2018) revealed that Indonesian Sign Language (BISINDO) arises naturally from the interaction of deaf people with their environment since childhood, which makes BISINDO also considered a deaf culture whose existence is felt to need to be raised to realize its existence and rights as a deaf person[1].

Anton Breva Yunada, et al (2018) concluded that a sign language recognition system using Microsoft Kinect can only produce an accuracy of 74% to 87.5% with a maximum detection distance of 150 cm for letter/word sign language recognition. [2]

Research by Prisky Ratna A. et al (2020) Revealed that creating an object detection application system that can detect electronic devices such as laptops, mice, cellphones, cameras, and also headphones by utilizing the SSD-MobileNet algorithm can produce an average of 93.02%[3].

Therefore, the main purpose of this proposed project is to build a deep learning-based model that can recognize the Indonesian cue-discussing alphabet (BISINDO) in complex environments.

2. Material and Methods

SSD-MobileNet is a pre-training object detection model that has been tested on the COCO dataset, so the weights in this model already exist based on previous training.

This model can be used with different objectives and only needs to fine-tune the short system to be created.

SSDs are one of the most popular single-stage object detectors used in some detection applications. Although the detection accuracy is not as good as the existing two-stage target detector, on the other hand, SSDs have the advantage of having a fast calculation speed. by using the VGG16 network as the backbone network model of the object detector. The VGG16 network provides six feature maps with different dimensions for back-end networks to detect multi-scale objects. Then, a non-maximum suppression (NMS) process is applied to the network model detection output. For each group with multiple overlapping detection outputs, the detection output with the highest confidence score is selected as the final detection result for that group. Although the VGG16 network model has good feature extraction capabilities, the network architecture is too large for an embedded platform. That makes VGG16 can exceed maximum system memory and it is difficult to achieve real-time performance when run on an embedded system.

To reduce the computational complexity of the VGG16-SSD detector, Google implemented the Mobilenet network model to replace the VGG16 network, improving the performance of the SSD detector in real-time. Figure 1 shows the existing MobileNet-SSD network architecture, which uses a second-generation Mobilenet network, called Mobilenet-v2, as the SSD detector backbone network model. The MobileNet-SSD detector inherits the VGG16-SSD design where the front-end Mobilenet-v2 network provides six feature maps of different dimensions for back-end detection networks to perform object detection at various scales. . Since the backbone network model was changed from VGG-16 to Mobilenet-v2, the MobileNet-SSD detector can achieve real-time performance and is faster than other existing object detection networks.

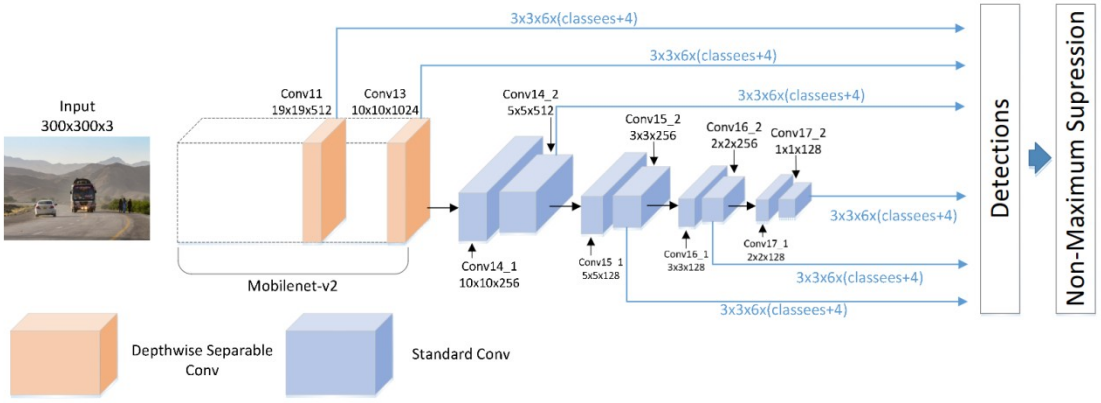

IMAGE 1. ARCHITECTURE SSD MOBILENET

The dataset used for the model training process is obtained from the Kaggle data center in the form of images with a total of 312 images with each letter totaling 12 images, the data set is then divided into 2 training data and also test data with a comparison of 70% training data and 30% testing data.

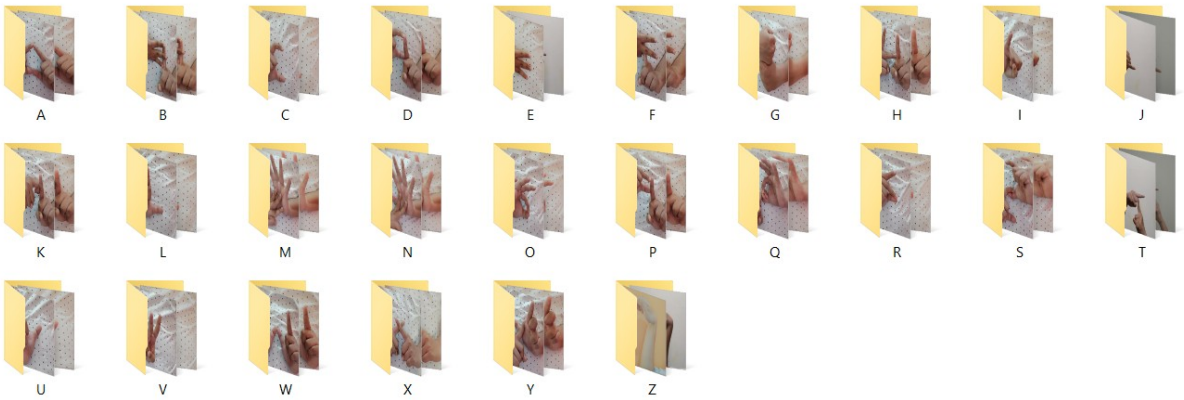

IMAGE 2. DATASET ALPHABET BISINDO

When run this system will check and ensure that there are no errors when the application is started, when there is no error the application will access the main camera on the device and will start reading the input generated by the camera and will be read every frame, which then the frame will be forwarded to be processed by the SSD-MobileNet model embedded in the system and when on the frame is detected the desired class it will be displayed to the inner screen containing information about what class was detected and also the accuracy generated at the time of detection, and if it is not detected it will again read the frame generated by the camera. More details can be seen in the following figure 3.

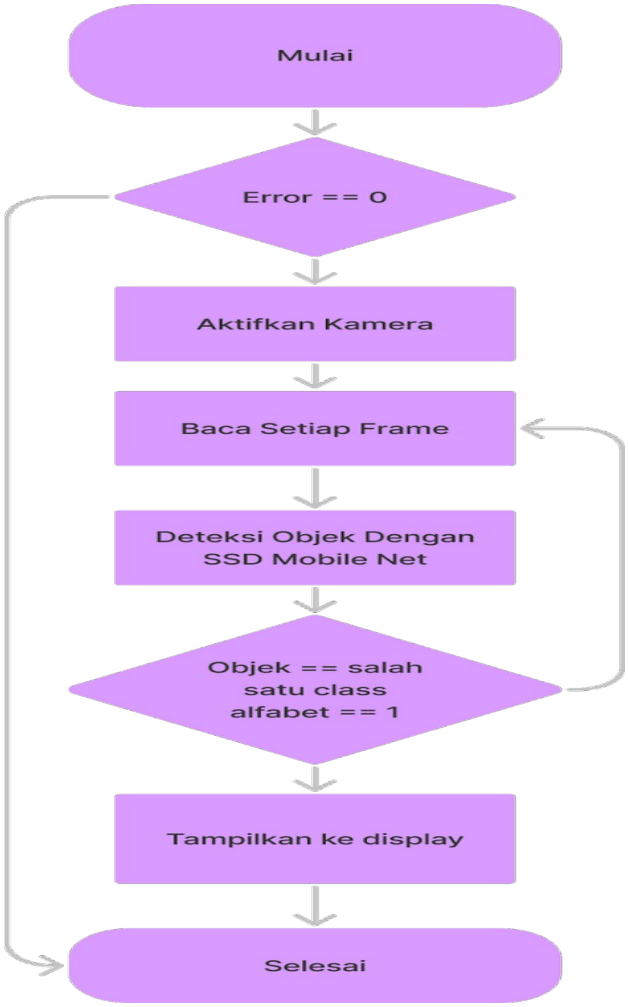

IMAGE 3. DIAGRAM FLOW CHART

3. Result and Discussion

The training process on the model was carried out using a data set with a size of 12 images for each class of 26 classes, training was carried out using the Windows 10 operating system using the Python programming language version 3.8 and also the Tensorflow 2 library version 2.9, using a computing system powered by an Intel(R) Core (TM) i7-9750H @ 2.60Ghz CPU, Memory 8GB@2666Mhz and also an Nvidia GeForce GTX 1650 4GB GPU.

The training process with a configuration of 50 epochs (the time all the data in the dataset was input into the neural network once) resulted in detection accuracy of more than 90%. More details can be seen in figure 4.

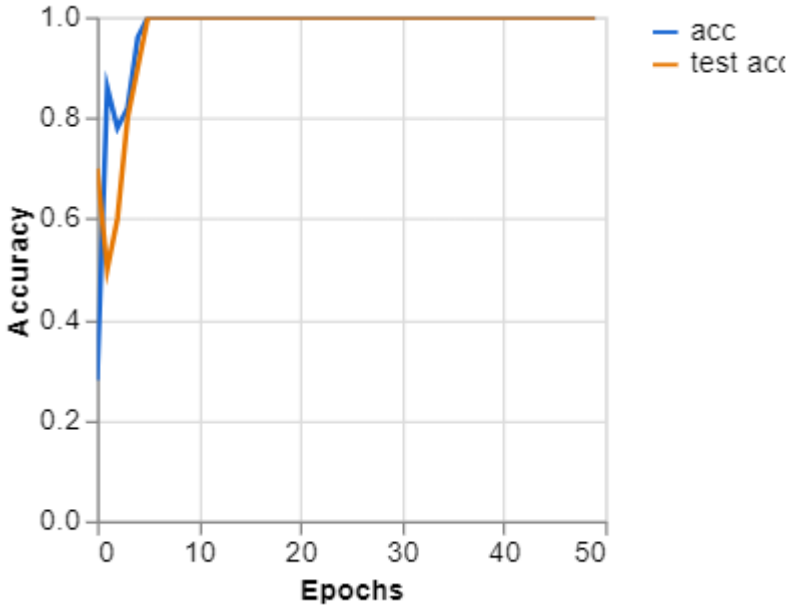

IMAGE 4. ACCURACY RESULTS IN THE TRAINING PROCESS

In figure 5, we can see that the loss rate of the function is at 1%, which means that the detection error made by this model is very small.

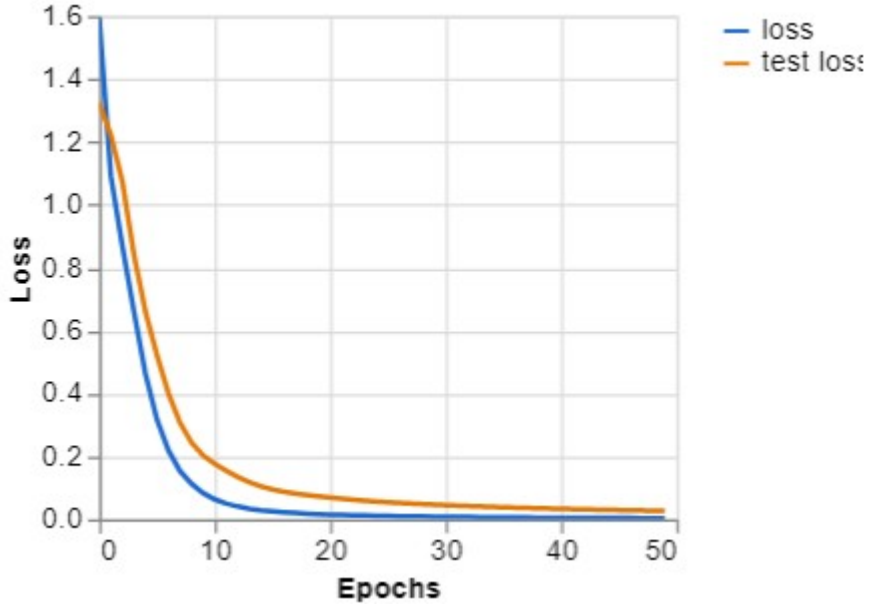

IMAGE 5. RESULTS OF LOSS IN THE TRAINING PROCESS

Application testing is carried out by using video that has been recorded with sufficient lighting conditions and input into the system and then the system will read and write a log of the results of the detection, this process is carried out 4 times repeatedly. The test results can be seen in table 1.

TABLE 1. TEST RESULTS

Symbol	Detection results	Accuracy (%)
	A	99.99768733978271
	B	99.97221827507019
	C	99.9931812286377
	D	99.99744892120361
	E	99.99678134918213
	F	99.98754262924194
	G	99.99518394470215
	H	99.95555281639099
	I	99.99583959579468
	J	99.9955415725708
	K	99.97729659080505
	L	99.97051358222961
	M	99.95050430297852

	N	99.97839331626892
	O	99.99488592147827
	P	99.99762773513794
	Q	99.98764991760254
	R	99.99861717224121
	S	99.99778270721436
	T	99.99803304672241
	U	99.98730421066284
	V	99.99774694442749
	W	99.99631643295288
	X	99.96877908706665
	Y	99.99659061431885
	Z	99.98123049736023

Based on the tests carried out, the average detection accuracy results were obtained as much as 99.98%, even so, this system can now only detect the alphabet and cannot recognize numbers, words, and even sentences, the system has also not been tested in various environmental circumstances so it still needs a lot of development and experimentation. The suggestion for the next research is to add the ability to the system to be able to detect words in BISINDO.

5. Reference list

[1] G. Gumelar, H. Hafiar, and P. Subekti, "THE CONSTRUCTION OF BISINDO'S MEANING AS A DEAF CULTURE FOR GERKATIN MEMBERS," *INFORMATION*, vol. 48, no. 1, p. 65, Jul. 2018, DOI: 10.21831/information.v48i1.17727.

[2] A. Breva Yunanda, F. Mandita, and A. Primasetya Armin, "Introduction to Indonesian Sign Language (BISINDO) For Character Letters Using Microsoft Kinect," *Fountain of Informatics Journal*, vol. 3, no. 2, p. 41, Nov. 2018, DOI: 10.21111/fij.v3i2.2469.

[3] Prisky Ratna A., Agus Sumin, and Setia Wirawan, "Making Object Detection Applications Using TensorFlow Object Detection API by Utilizing V2 MobileNet SSDs as Pre-Trained Models," *Scientific Journal of Computing*, vol. 19, no. 3, Mar. 2020, DOI: 10.32409/jikstik.19.3.68.