Vision-Transformer on Unconstrained Ear Recognition (ViT-Ear)

Marwin B. Alejo
2020-20221
marwin.alejo@eee.upd.edu.ph

I. Brief Introduction and Motivation

- Ear recognition is a subset of object detection with ear images containing a unique human identity signature as the subject for detection.
- Unconstrained ear recognition as a subset of ear recognition aims to recognize human identities through ear images taken from the wild. It had become a budding and open sector of biometric research in recent years, imploring computer vision methods.
- Past studies utilize image processing algorithms for the fabrication of an unconstrained ear recognition system. However, it is tedious and requires a large number of computing resources. [1, 2, 3]
- Recent studies suggested using different hand-crafted CNN architectures and algorithms to fabricate an unconstrained ear recognition system. However, the limited and small number of unconstrained ear datasets hinders the efficient implementation of these CNN algorithms as it requires a large number of ear images. [1, 2, 3]
- As a resort to the limitations of the recent studies, the concept of transfer learning is introduced to unconstrained ear recognition to extend CNN with only a limited number of unconstrained ear images without relying on hand-crafted and complex algorithms for practical and efficient system development. [1, 2, 3, 4]
- Given the above limitations and outcomes of former unconstrained ear recognition studies and the field itself, this mini-project generally aims to extend a pure Transformer Network to the task of unconstrained ear recognition. Transformer in deep learning is a nascent technology used initially in NLP but recently attracts researchers to explore and extend it in Computer Vision applications.

II. Objectives

This mini-project aims to extend vanilla Transformer Network to Unconstrained Ear Recognition as a computer vision task and a subset of object detection with the following specific goals:

1. Measure the transformer-based unconstrained ear recognition network's performance in terms of validation loss and accuracy; and

2. Compare the validation loss and accuracy as performance metrics of the Transformer-based network with CNN-based through transfer learning network for unconstrained ear recognition.

III. Limitations of the Mini-project

As a mini-project, this study is limited only to the exploration of Transformer Network, specifically Vision-Transformer (ViT) without the inclusion of hybrid networks, with unconstrained ear recognition by extension and compare its performance with the performance of the published results of the same task but uses a CNN-based approach. Furthermore, considering constraints such as time and computing resources (Google Colab Free), this mini-project does not aim to develop an unconstrained ear recognition prototype or an end-to-end system.

IV. Review of Related Works

Transformers first appeared as a simple and scalable solution to attain state-of-the-art results in NLP and is currently being extended in Computer Vision. Like transfer learning in CNN, Transformer Networks may be trained over large datasets and fine-tuned to learn on small datasets. Among the most recently developed transformer architectures for computer vision are Detection-Transformer (DeTr) [5] for and Vision-Transformer (ViT) [6]. While both achieved state-of-the-art results, only Vision-Transformer (ViT) is possible for recognition tasks as DeTr is limited with detection tasks.

There have been no published studies relevant to any transformer-based recognition task at the moment of this writing except for Vision-Transformer as is. Most literature works are subjected to either complex hand-crafted development of a Transformer Network or in the extension of detection tasks.

The Vision-Transformer

Vision-Transformer (ViT) employs a modified transformer network architecture as such it may directly operate on images instead of text. ViT divides an image into grid square patches. Each patch is flattened into a single vector by joining all the channels of pixels in a patch and linearly inject it into the desired dimension. For it to learn, a learnable position embedding is added to each patch and allows the Transformer Network to learn the image's positional patch. Figure 1 shows the ViT architecture as shown in its original paper.

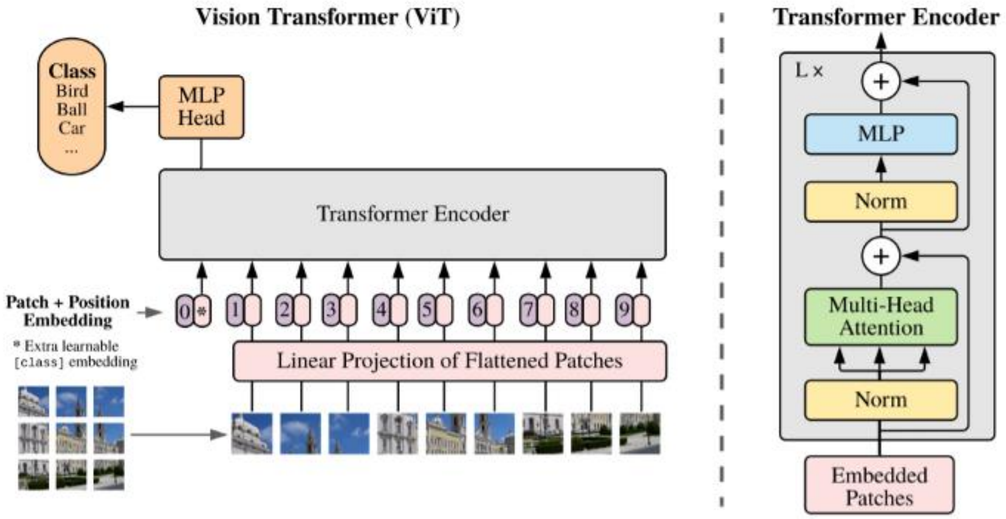

Figure 1 Vision-Transformer (ViT) Architecture [6]

V. Methodological Approach

Figure 2 shows the method used in this mini-project in chronological order.

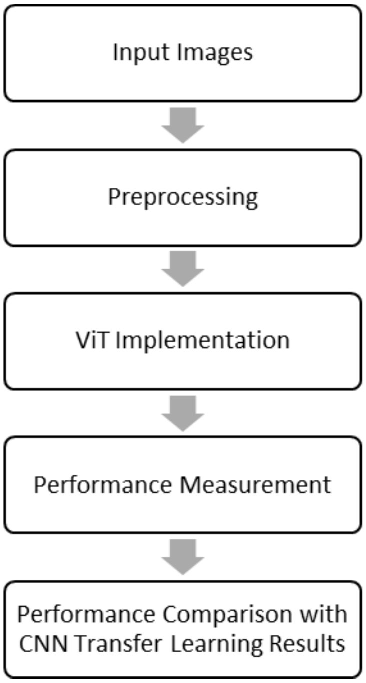

Figure 2 Methods used in the mini-project.

Input Images

EarVN1.0 [7] is the used dataset in this mini-project. This dataset consists of unprocessed unconstrained ear images of 164 people, with each having ~180 images, for a total of 28,412 ear images. However, in consideration of the computing resources constraint for the unconstrained ear recognition task, only the first 20 classes of the EarVN dataset is used for a total of ~4000 images. Figure 03 shows a sample of the used ear images in this mini-project.

Images from each classes is split into training set with 80% of the total images and testing/validation set with 20% of the total images.

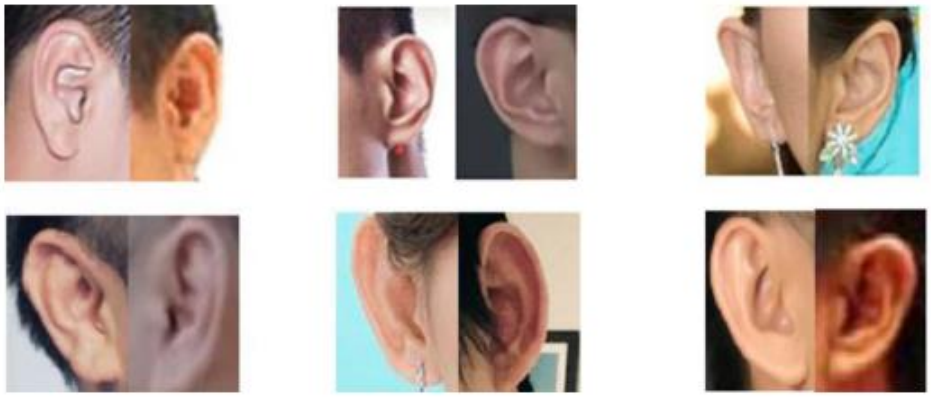

Figure 3 Sample ear images of EarVN1.0

Preprocessing

Preprocessing used in this mini-project includes image resizing to 224 square pixels, horizontal and vertical flipping of an image to 30 degrees, and the standard ImageNet normalization values. These preprocessing steps are applied to the training set while only the resizing and image normalization for testing/validation images

ViT Implementation

Following the ViT description, ViT for unconstrained ear recognition is implemented with the pre-trained ImageNet-21k model and fine-tuned to classify or recognize 20 people. Vision-Transformer is implemented in this mini-project using Pytorch with XLA on Google Colab TPU. The model is trained on 20 epochs with the following configuration as shown in table 1. Overfitting occurs on models trained beyond 20 epochs.

Table 1 ViT Implementation configuration

Training Parameter	Value
Batch size	8
Learning Rate	0.00002 on Adam
Gamma	0.7
Epochs	20

ViT Performance and Results

Vision-Transformer on Unconstrained Ear Recognition achieves a validation accuracy of 95.31% with a loss of 26.36%. This outcome is closely comparable to the CNN-based transfer learning results on the same task [4]. Table 2 shows the comparative results of ViT on Unconstrained Ear Recognition and the accuracy results of state-of-the-art CNN architectures on the same task with transfer learning *(from my previous study).*

Architecture	Accuracy (%)
This mini-project (ViT-Ear)	**95.31**
ResNet18 (Transfer Learning)	100.00
AlexNet (Transfer Learning)	97.30
ResNet50 (Transfer Learning)	96.70
Inception-ResNet (Transfer Learning)	94.70
Inception-v3 (Transfer Learning)	96.70
GoogLeNet (Transfer Learning)	93.30
SqueezeNet (Transfer Learning)	87.30
ShuffleNet (Transfer Learning)	86.70
MobileNetv2 (Transfer Learning)	81.30

VI. Conclusion

In this mini-project, Transformer Network is extended to unconstrained ear recognition with Vision-Transformer. Vision-Transformer achieves an accuracy result that is closely comparable to the state-of-the-art CNN-models' accuracy results on the same task.

VII. Recommendations

Since this is a mini-study and shortly time-bounded, it is recommended to extend this into an end-to-end system incorporating Detection-Transformer (DeTr) and Vision-Transformer (ViT) as one. Also, it is recommended to use another fabricated unconstrained ear dataset similar to the make of Annotated Web Ears (AWE) of UERC.

VIII. References

[1] Ž. Emeršič, V. Štruc and P. Peer, "Ear Recognition: More Than a Survey," *CoRR,* vol. abs/1611.06203, 2016.

[2] Ž. Emeršič, D. Štepec, V. Štruc and P. Peer, "Training Convolutional Neural Networks with Limited Training Data for Ear Recognition in the Wild," in *IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, Washington, DC, 2017.

[3] E. Z. e. Al., "The Unconstrained Ear Recognition Challenge 2019," *CoRR,* 2019.

[4] M. Alejo and C. P. Hate, "Unconstrained Ear Recognition through Domain Adaptive Deep Learning Models of Convolutional Neural Network," *International Journal or Recent Technology and Engineering,* vol. 8, no. 2, pp. 3143-3150, 2019.

[5] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov and S. Zagoruyko, "End-to-End Object Detection with Transformers," *Lecture Notes in Computer Science,* p. 213–229, May 2020.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *CoRR,* October 2020.

[7] V. T. Hoang, "EarVN1.0: A new large-scale ear images dataset in the wild," *Data in Brief,* vol. 27, 2019.