



Machine-learning Approach for Robot Tactile Representation and Perception

Final Year Project: Interim Report

Ouyang Yanjia

Supervisors: Assoc. Prof. Lin Zhiping, Dr. Wu Yan

School of Electrical and Electronic Engineering Nanyang

Technological University

November 2021

Contents

1	Introduction	2
1.1	Tactile Dataset	2
1.2	Previous Methods.....	3
2	Current Progress	3
2.1	Transformer Model	3
2.1.1	Transformer Architecture	3
2.1.2	Attention Mechanism	4
2.1.3	Vision Transformer (ViT).....	5
2.1.4	Model Parameters	5
2.1.5	Results	6
2.2	Time-Space Transformer Model	6
2.2.1	Model Structure	6
2.2.2	Results	7
2.2.3	50 Classes Dataset.....	8
3	Future Plan.....	8
3.1	Model Compression	8
3.2	Final Report.....	9
4	Conclusion	9

1 Introduction

With the development of technology, people have great expectations for robots that not only can handle specific tasks in fixed stations but be able to perceive the environment and interact with human beings. Such kind of intelligence depends on the ability to fuse and process sensors data. In this project, we focus on tactile data of different materials and use advanced machine learning techniques to do texture classification.

1.1 Tactile Dataset

Tactile data is provided by tactile sensors which aim to simulate the action of human touch. Those sensors are mostly bionic based and the mechanism behind various sensors may deviate from sensing pressure to measuring temperature. Two sets of data are used in this project [1]. One is from iCub robot (XI dataset) and the other is from BioTac robot (XB dataset). Each dataset is collected by sliding the sensor over 20 different materials which are shown in Figure below. 50 samples are included for each class.

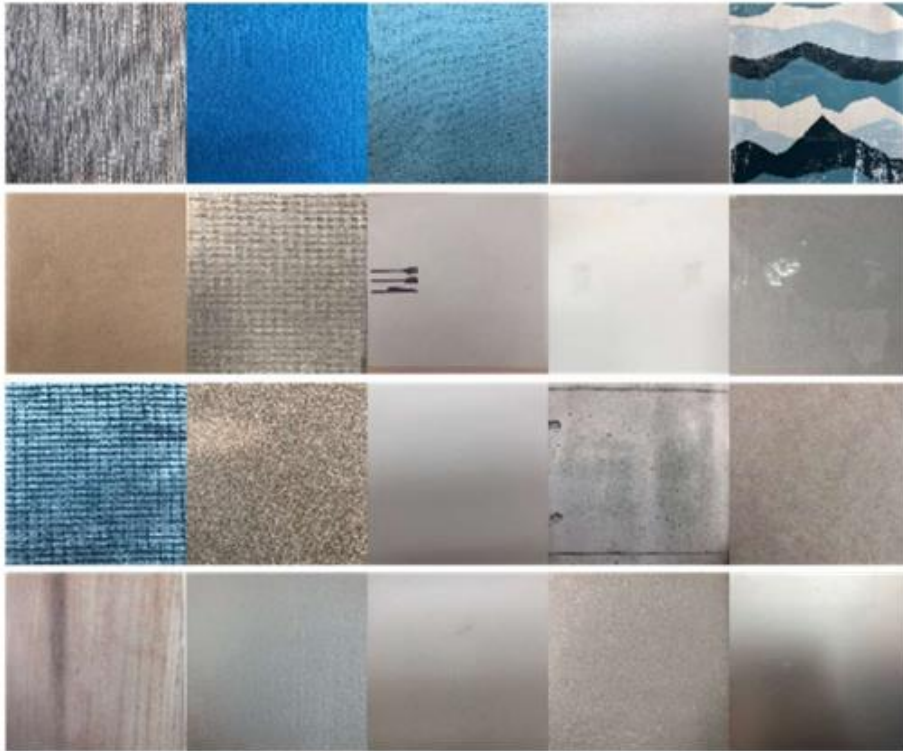


Figure 1. Snapshots of 20 materials [2]

As the sensor slides over the material, sequential measurements are recorded at each time step, making the data have both spatial and temporal context. This feature implies the model needs to consider both contexts to achieve a relatively good result.

1.2 Previous Methods

To efficiently classify the tactile dataset, different models have been proposed. One model is built on the spiking neural network (SNN) structure [3].

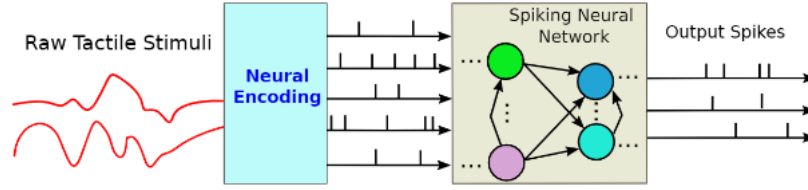


Figure 2 SNN model framework [3]

Another model is based on recurrent autoencoder formed by LSTM block with a header network [1].

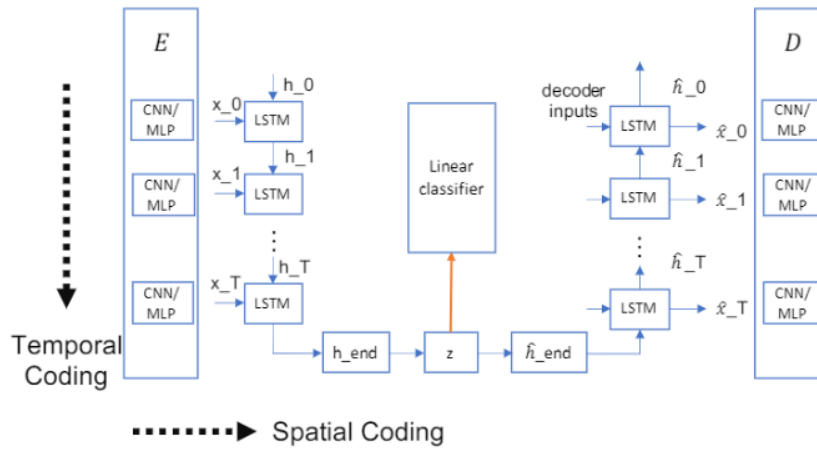


Figure 3 LSTM model framework [1]

2 Current Progress

2.1 Transformer Model

2.1.1 Transformer Architecture

Transformer architecture has shown great capability in natural language processing tasks [4]. It uses purely attention mechanism to build encoder and decoder layers. The attention mechanism, also called self-attention layer, calculates the attention which reveals the similarity or correlation between each part of data. Therefore, the model is able to use this attention to relate input and output by merging global information.

Before the encoder-decoder block, inputs are cut into small pieces and “position encodings” are added to each piece. This token is aimed to reserve the relative or absolute position information of the original data. The position encoding value can be either learned in training or fixed by predefined equations.

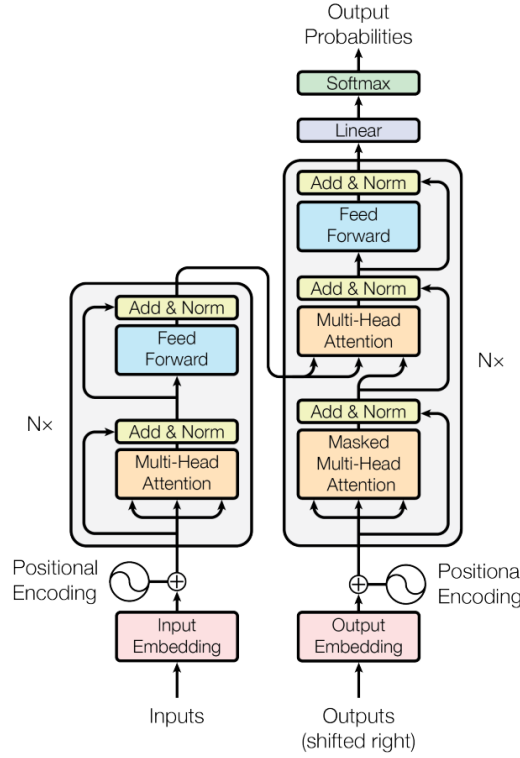


Figure 4 Transformer model architecture [4]

2.1.2 Attention Mechanism

During input embedding, long input will be cut into short pieces. Each piece will have three matrices with the same dimension but different weights. The three matrices are matrix Q, K and V, which stands for query, key and value. The attention is computed using three sets of matrices:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

$\frac{1}{\sqrt{d_k}}$ is just a scaling factor to prevent gradient being too small

For every piece, it will map its query matrix to key matrix and value matrix of every matrix (including itself). Then, each computed attention will be added together as the output of the specific piece.

Multi-Head Attention is the enhancement of attention. It enables the model to gain different aspects of information to achieve better results.

$$MultiHead(Q, K, V) = Concat(head_1, \dots, head_h)W^O$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$

For classification tasks, after Multi-Head Attention, a class token will be added. This token will be used for the final determination of the class prediction.

2.1.3 Vision Transformer (ViT)

Vision Transformer inherited the basic structure of transformer but implemented different embedding schemes [5]. In simple terms, it cuts the image into small patches and then resizes each patch into one dimension vector so that the original transformer structure can be used.

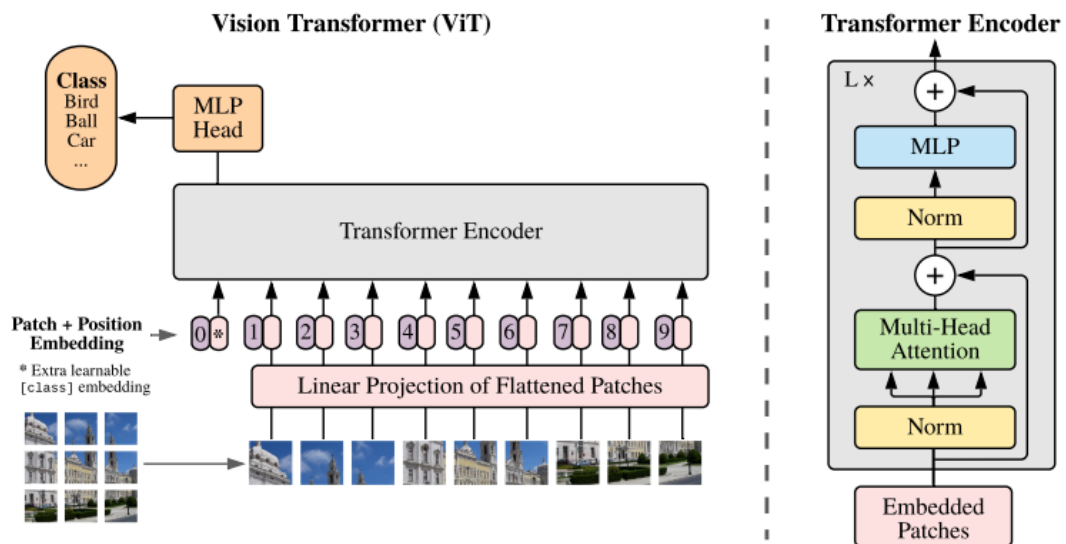


Figure 5 Vision transformer overview [5]

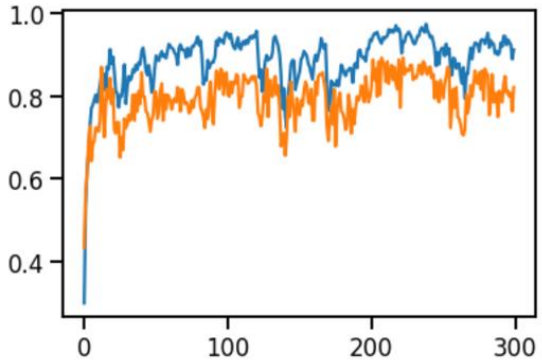
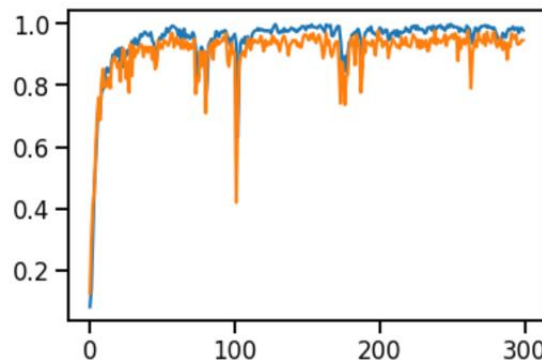
2.1.4 Model Parameters

In XI dataset, sensor outputs form a 6×10 size matrix at each time-step with length 75. And the XB dataset is 1×19 size with 400 time-steps for one sample. Since the ViT model only takes in 2-D images, I concatenate all sensor outputs into one big “image”. Therefore, the samples size for XI dataset is 6×750 and for XB dataset is 19×400 . In this way, the temporal information of each sensor output (from start to the end) is converted into spatial information (from left to right) of the concatenated “image”.

Then, for XI dataset, the input image size is set to (6,750), and the patch size is (3,5). For XB dataset, the input image size is (19,400), and the patch size is (19,40). There are in total 20 materials to be classified so the number of classes is 20 and the loss function is cross-entropy loss. Adam is used as optimization function, and the learning rate starts at 0.001 and decreases with each epoch.

2.1.5 Results

The ViT model gives out good results in two datasets, especially for the XB dataset. The main reason for the difference in accuracy is that the XB dataset has lower noise. Such a clean dataset makes the model easy to find the latent pattern between inputs and outputs, while the noisy data may face some level of overfitting problems.

Dataset	Best validation accuracy	Train accuracy (blue) and validation accuracy (orange) in 300 epochs
XI	0.8884	
XB	0.9777	

2.2 Time-Space Transformer Model

2.2.1 Model Structure

Many researchers have tried to apply transformer model into video recognition field [6][7]. Compare to image alone, video contains both spatial and temporal information. Thus, redesigning the transformer structure to adapt to the two dimensions may lead to better performance. Several space-time structures are proposed and implemented as ViViT model [6]. Here, an intuition one is chosen which is shown on the left in figure 6.

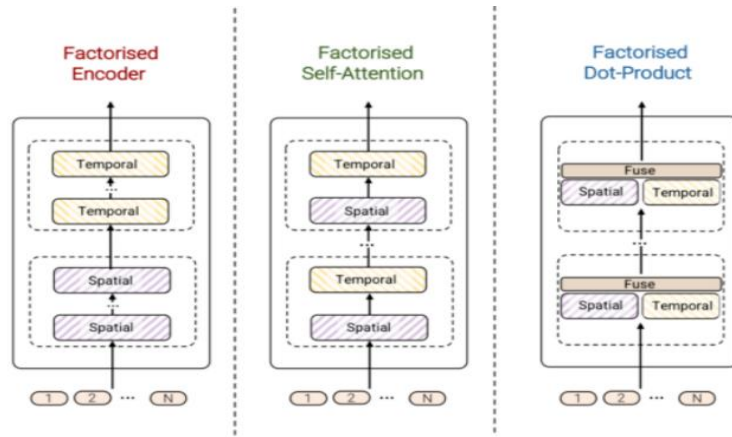


Figure 6 Space-time Structures

The model consists of two separate transformer encoders. The spatial encoder in the first part only exchanges the tokens extracted from the same frame, and then the time encoder exchanges the tokens at the uniform position of different frames, and finally is classified by a multilayer perceptron (MLP).

2.2.2 Results

Dataset	Best validation accuracy	Train accuracy (blue) and validation accuracy (orange) in 300 epochs
XI	0.9598	
XB	0.9911	

Compare to the previous ViT model, this space-time model obviously has better performance in both accuracy and stability. For XI data, the gap between the train set and the validation set is smaller which means this model is more robust

and reduce overfit problem.

Moreover, the parameter efficiency is also improved. Although this model has more transformer layers, it has fewer inputs for each layer, therefore, the total trainable parameters is much fewer. Ignoring the slight difference between two dataset, ViT model needs approximately 11.226M parameters, while this model only needs about 0.158M parameters.

2.2.3 50 Classes Dataset

The model is then applied to a new XB dataset with 50 classes, with 50 samples in each class, whose previously accuracy is 0.95 [2]. The best validation accuracy using this spatial-temporal transformer model is 0.9648, and the accuracy in the test dataset is 0.9617.

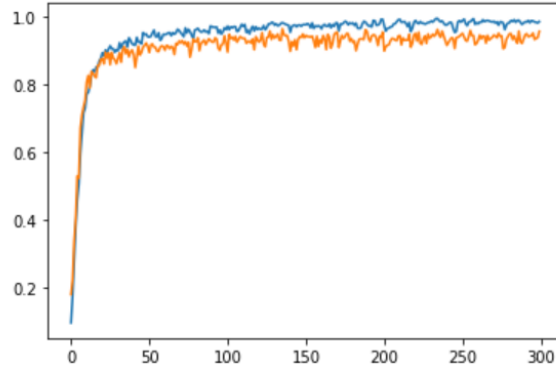


Figure 7 Learning Curve of 50-class XB Dataset

3 Future Plan

3.1 Model Compression

Although the current model is about 100 times smaller and gets better accuracy compared to the original ViT model, it can still be improved since the LSTM model uses even fewer parameters [1]. For example, the VTN model uses “Longformer” structure to adapt long sequences which also reduced model size. Other compression methods like K-means quantization or binarization, low-rank factorization and knowledge distillation may help to reduce the size and squeeze the potential of the transformer model.

Estimated Timeline: 1st Dec - 17th Dec

3.2 Final Report

If the improved transformer model has very good performance, the final report can start early and there will be more time to deal with possible problems or investigate possible new ideas. Also, there will be time to organize the code and documentation.

Estimated Timeline: 20th Dec - 28th Jan

4 Conclusion

Up to now, all the tasks are in progress in an orderly manner. Although the Covid-19 situation affects the routine that going to the lab is not that easy, the improvement of the model is still proceeding and communication with professors is still smooth. Up to now, the transformer model has already shown great ability in this task, which makes us believe its potential in further exploration. Hope we can find more interesting things in digging into the transformer model.

Reference

- [1] R. Gao, T. Taunyazov, Z. Lin, and Y. Wu, "Supervised autoencoder joint learning on heterogeneous tactile sensory data: Improving material classification performance," 2020 presented.
- [2] R. Gao, T. Tian, Z. Lin, and Y. Wu, "On Explainability and Sensor-Transferability of a Robot Tactile Texture Representation Using a Two-Stage Recurrent Networks ," 2021.
- [3] T. Taunyazov, Y. Chua, R. Gao, H. Soh and Y. Wu. "Fast Texture Classification Using Tactile Neural Coding and Spiking Neural Network." 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2020): 9890-9895.
- [4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv.org*, 06-Dec-2017. [Online]. Available: <https://arxiv.org/abs/1706.03762>. [Accessed: 15-Nov-2021].
- [5] 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," *arXiv.org*, 03-Jun-2021. [Online]. Available: <https://arxiv.org/abs/2010.11929>. [Accessed: 15-Nov-2021].
- [6] A. Arnab, C. Schmid, M. Lucic, C. Sun, G. Heigold, and M. Dehghani, "Vivit: A video vision transformer - arxiv." [Online]. Available: <https://arxiv.org/pdf/2103.15691.pdf>. [Accessed: 17-Nov-2021].
- [7] D. Neimark, O. Bar, M. Zohar, and D. Asselmann, "Video Transformer Network," *arXiv.org*, 17-Aug-2021. [Online]. Available: <https://arxiv.org/abs/2102.00719>. [Accessed: 17-Nov-2021].