Design and Implementation of Deep Convolutional Siamese Network for Image Recognition Based on One-Shot-Learning

Tang Chi-En (湯琦恩)

Abstract

In today's society, Artificial Intelligence for facial recognition have become indispensable in various aspects of our lives, one network architecture called **Siamese Network** plays a crucial role in this kind of tasks. However, the machine learning often require a large amount of data and can be very computationally expensive. But in certain practical scenarios, obtaining a sufficient amount of training data is not always feasible. Therefore, it would be convenient and efficient to train models using fewer training data. This is where **meta-learning** comes into play.

In this study, I proposed a Siamese network architecture and explore more effective loss functions based on several commonly used ones. The research results demonstrate that by using the adapted base network, contrastive loss function, Adam optimizer, and ReLU activation function, remarkable recognition accuracies of 97.27%, 76%, and 75.63% are achieved on test data from the MNIST, LFW, and AT&T datasets, respectively, with a small amount of data (based on the One-Shot learning framework) and fewer network parameters. Additionally, the training convergence speed is faster (compare with the Siamese Neural Networks for One-shot Image Recognition study [4]) .

I. Introduction

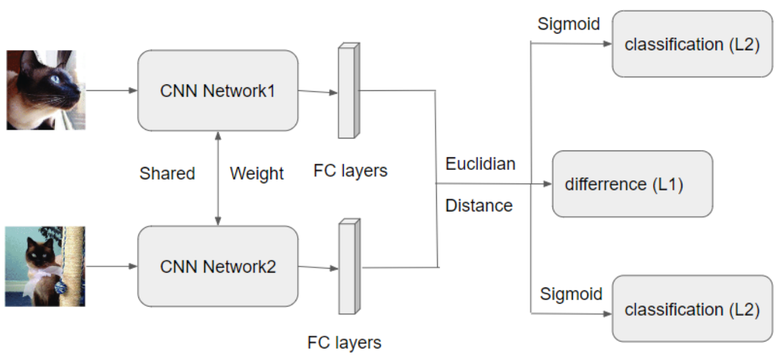
While meta-learning is indeed an effective learning paradigm for limited amount of training data, it remains challenging to achieve the same level of accuracy as networks trained on massive datasets.

Moreover, in practical applications, the use of large-scale training data is still prevalent, with meta-learning frameworks primarily employed for evaluating network performance.

Therefore, this study aims to design a training practices that utilizes a small amount of input data (based on the meta-learning framework). The training protocol will encompass the selection of Siamese network architecture, optimizer, activation function, loss function, input data format, as well as settings for learning rate and initialization. The objective is to verify whether this approach can improve recognition accuracy and model generalization, even with insufficient input data.

II. Material and Method

2.1 Model Design

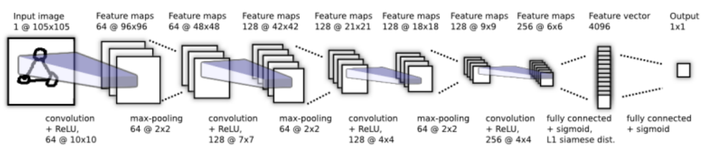


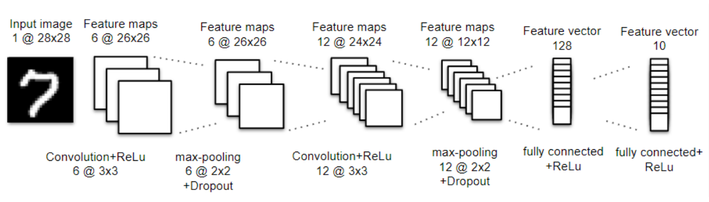
- **Siamese Networks**：The network architecture consists of two neural networks with shared weights and identical structures. It ranks similarity between two inputs. The objective of the Siamese network is to minimize the distance between similar images or bring them closer together in the feature space.

2.2 Architecture of Neural Network

- **Different Architecture of Siamese Network Using in different datasets:**

① LFW (Focal Loss)	Total params: 27,417,409
① MNIST ② AT&T ③ LFW (Contrastive Loss)	Total params: 40,538





2.3 Settings of training

- **Optimizer**：Compare performance from Adam、RMSprop、SGD, learning rate sets 5e-4.
- **Initialize**：Network initialize follows normal distribution and weight regularization is applied in convolutional layers using L2 (penalty term) to prevent overfitting.
- **Input Data type**：Different Loss function has differ type of input data. In Focal Loss using 10-Way One-Shot support set；Contrastive Loss uses 200 training pairs.
- LFW dataset has multiple classes and can be trained and tested using the One-Shot learning framework with a total of 5749 images. Other datasets only utilize One-Shot learning for the testing task (10-Way One-Shot support set).
- **Active Function**：Compare performance from ReLU、Sigmoid、Softmax.
- **Data Augmentation**：For each class in the LFW dataset, only one image is taken. The second image required for the training pair input of the Siamese network is generated by data augmentation such as affine transformation or clockwise (counterclockwise) rotation .

2.4 Loss Function

- **L1-similarity based Loss function：Contrastive Loss**

$$L_1 = L(W,Y,X_1,X_2) = (1-Y)\frac{1}{2}(D_W)^2 + \frac{1}{2}Y\{\max(0, m - D_W)\}^2 \qquad D_W = ||X_1 - X_2||^2$$

- **L2-multiple classification Loss function： Focal Loss**

$$L_f = (-\alpha_t(1-p_t)^\gamma \log(p_t)) \ , \ p_t = \begin{cases} p & y = 1 \\ 1-p & else \end{cases}$$

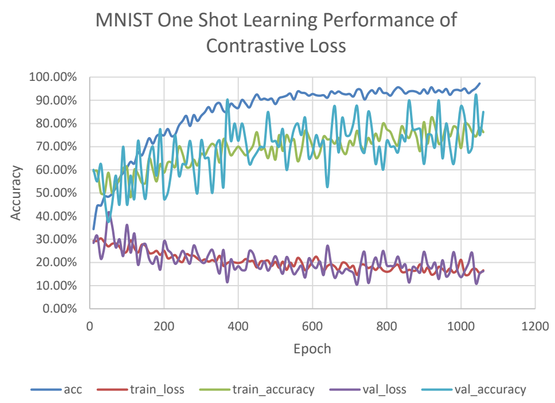
- **L2-multiple classification Loss function： L-Softmax Loss**

$$L_s = \left(-log\frac{e^{\|W_{yi}\|\|X_i\|\varphi(\theta_{yi})}}{e^{\|W_{yi}\|\|X_i\|\varphi(\theta_{yi})} + \sum_{j\neq yi}\left(e^{\|W_j\|\|X_i\|\varphi(\theta_j)}\right)}\right) \qquad \varphi(\theta) = \begin{cases} \cos(m\theta) & 0 \le \theta \le \frac{\pi}{m} \\ D(\theta) & \frac{\pi}{m} \le \theta \le \pi \end{cases}$$

III. Result

- **MNIST Dataset**

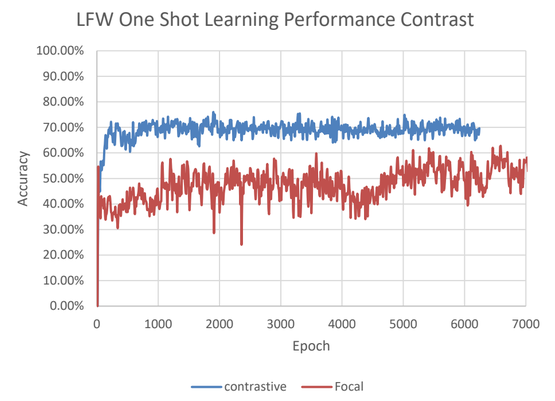
MNIST test accuracy	
Model	Accuracy
Humans	95.5
Hierarchical Bayesian Program Learning	95.2
Deep Boltzmann Machine	62.0
Siamese Neural Network	70.3
Siamese-Focal Loss	**91.0**
Siamese-Contrastive Loss	**97.27**



- Test accuracy：Achieved **97.27%** is higher compared to other Siamese network architectures that also utilize meta-learning, which achieved an accuracy of **70.3%**.
- Convergence speed：The Siamese network using Contrastive Loss achieved faster convergence speed (**1060 epochs**) compared to using Focal Loss or other network architectures, which required a significantly higher number of epochs (**900,000 epochs**).
- Total params：The network architecture used for training had significantly fewer parameters (**40,538**) compared to other network architectures (**38,954,049**).

- **LFW Dataset**

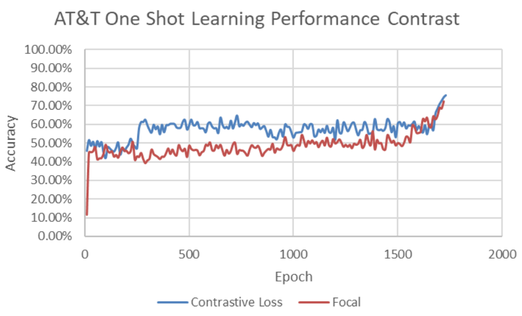
LFW		
Model	Siamese Neural Network	**Siamese-Contrastive Loss**
Accuracy	68.59	**76.0**
Epochs	500,000	**6240**
Params	38,954,049	**40,538**
Data	13,233	**11,498**



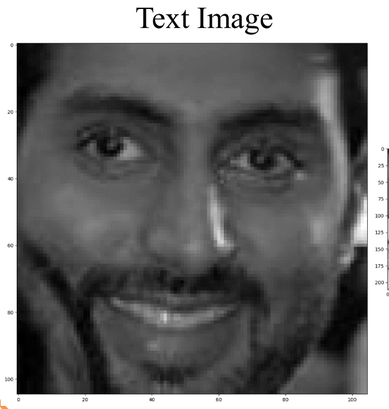
- Amount of data：The amount of data used for training and testing (including both original data of **5749** images and augmented data of 11,498 images) is significantly lower compared to other training methods (13,233 images).
- This implement is achieved through the utilization of the **One-Shot learning** framework.
- Convergence speed： Using Contrastive Loss achieved faster convergence speed (**6240 epochs**) compared to other network architectures, which required a significantly higher number of epochs (**500,000 epochs**).
- Total params： The network architecture used for training had significantly fewer parameters (**40,538**) compared to other network architectures (**38,954,049**).

- **AT&T Dataset**

AT&T		
Model	Siamese Neural Network	**Siamese-Contrastive Loss**
Accuracy	75.0	**75.63**
Params	4,018,650	**40,538**
Data	400	**400**

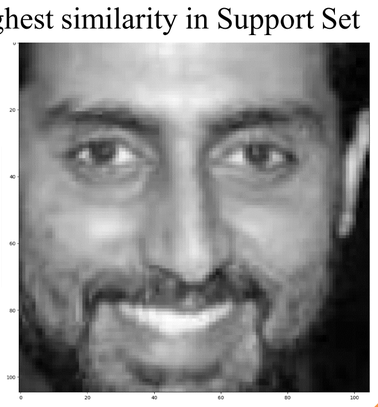


Text Image



Support Set

Highest similarity in Support Set



IV. Conclusion

The current implementation has successfully achieved the One-Shot learning framework, **reducing training costs** by lowering the number of network parameters and using a small amount of training data. It has achieved an accuracy of 97.27% in handwritten character recognition and 76% accuracy in more complex face datasets. The next step is to explore other optimization methods to further enhance accuracy and improve the model's generalization capabilities across different datasets. Additionally, efforts will be made to improve the overfitting issue with L-Softmax Loss.