Final Project Proposal

Team 15 Member:

Proposal format:

-Milk yield prediction

-Methods

1. Data preprocessing

(1) Label encoding for label type features (ex. 精液編號、酪農場代號…), and normalized the continuous ones.

(2) Dimension reduction:

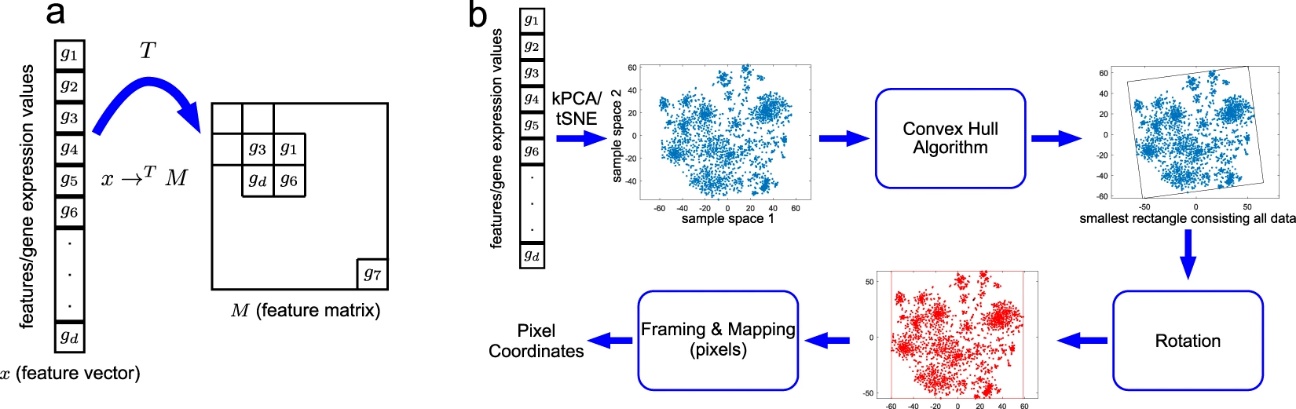
Since there is a large amount of feature after label encoding, it is possible that the input data is sparse. Dimension reduction could help extracting the efficient features. Here, classification of the labels may be high dimension (ex. 201 classes of 最後配種精液 in report.csv), so t-distributed stochastic neighbor embedding is recommended due to its relatively small reconstruction loss.

(3) Convert features into image:

[Reference] DeepInsight: A methodology to transform a non-image data to an image for convolution neural network architecture (<https://www.nature.com/articles/s41598-019-47765-6>)

In order to apply ResNet into non-image dataset, we have to design a kind of way converting source dataset into image-like dataset. After doing some literature review, we found a methodology, DeepInsight, to transform non-image data to image from Scientific Reports. The general concept of this method is putting the features that are highly “similar” to each other as neighbors on the image. (See the left part of figure below)

Recently, the most popular methods to measure the similarity between features within dataset are t-distributed stochastic neighbor embedding (t-SNE) and kernel principle component analysis (kPCA). We choose to apply t-SNE to dataset as the same method used in the essay. After t-SNE processing, we could get a 2D plane with lots of points that represent the location of features with respect to their similarities. Next we apply the convex hull algorithm to find the smallest rectangle containing all the points. Last but not least, because the image should be framed in a horizontal or vertical, we perform rotations to the convex hull, and finally get images which indicate transformation information from non-image data to image-like data. (See the right part of figure below)



2. Model

1. ResNet framework:

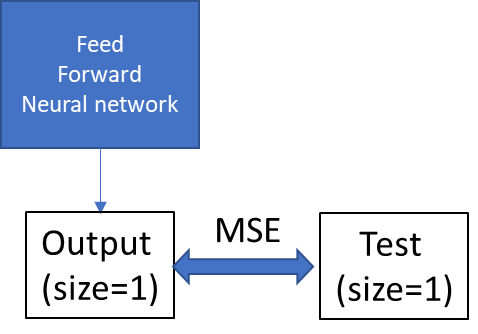
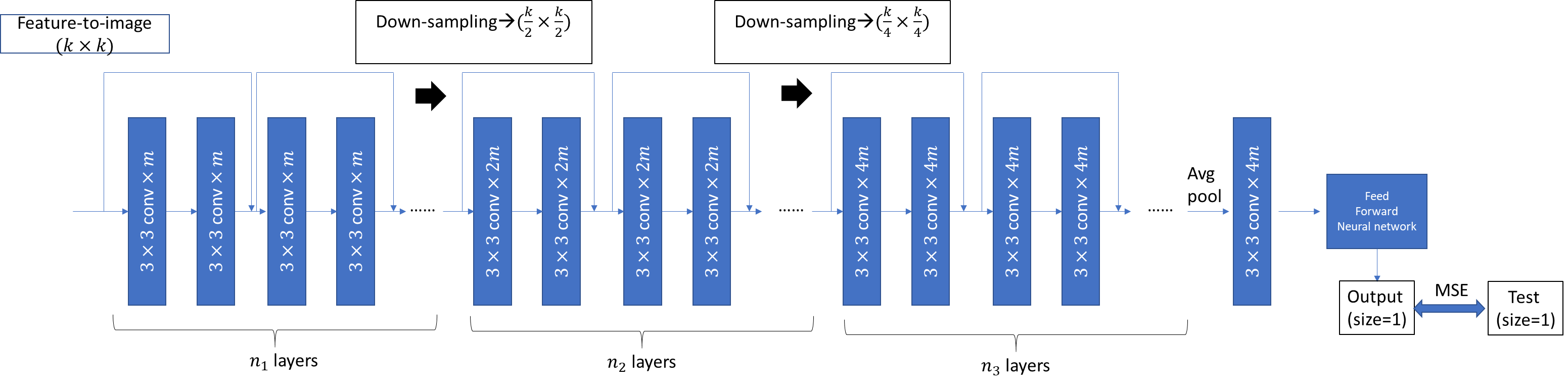
Introducing ResNet into model has been successfully confirmed to take image classification problem into another level via the residual function and shortcut connection. Therefore, apply the network on the image-like input data is inferred to enhance model performance. The framework is showed in figure.1.

1. Feed forward neural network:

After the residual network and down-sampling (average pooling), the high dimensional feature map () enters a feed-forward neural network, and returns an output with size 1.

1. Prediction:

Mean squared error is chosen as the loss function according to the regression problem. Also, applying ensemble methods to optimize the model performance by aggregating 3~5 candidate models (e.g. averaged), it could result a more comprehensive model.



**Reference**

附件X是關於Final project proposal的參考範本，一組一個代表繳交即可。

本次作業會算分，請同學在撰寫proposal時要列出欲使用明確的方法，最好可以畫flowchart整理流程，請不要呈現太簡陋。

The attached file is the template of final project proposal. Only one file needs to be uploaded for each team.

This homework will be considered for grading. We recommend drawing a flowchart to represent the training and predicting process. Please present the method you are going to use and write clearly.