# Image selection based on autoencoder neural network and application to the semi-supervised image classification

Tushar Singh[*1] Ashish Kumar Gaurav[*1*2] Yasuhiro Tsuchida[*1] Fadoua Ghourabi[*1]

[*1] AWL Inc., Japan [*2] IIT Kharagpur, India

Convolutional neural networks (CNNs) are becoming a key technology in processing and analyzing real-time video streams, such as security videos. When pre-processing video streams for training CNNs by splitting into image frames, we generate a large-scale image dataset from which a subset is used for training models. The random selection of a subset ignores the properties of the data and produces a repetitive dataset, which is not useful for training. This paper presents an image selection approach based on the autoencoder neural network. The autoencoder projects high-dimensional image feature vectors into a low-dimension latent space for effective analysis of image similarity. This approach allows not only to select representative images but also to facilitate the pseudo-labeling of unlabeled data. In this paper, through experiments with autoencoder, we show the benefits of this method in selecting images for training. We also explain the application to a semi-supervised image classification problem where our approach significantly enhances the accuracy comparing to random selection.

## 1. Introduction

Deep learning-based computer vision has strengthened the real-time monitoring of security videos. With capabilities of face recognition, tracking, classification of objects, AI applications such as theft prevention are emerging. This research, for instance, is motivated by providing reliable AI application for theft prevention to retail companies in Japan. AI theft prevention requires, as the first step, to detect persons, then classify them as *customer* or *staff*. Next, tracking algorithms and action recognition are run to detect and notify suspicious behavior. In this paper, we present our method for developing customer/staff classification model for processing security videos. The classification is a classical application of a common neural network for image processing, known as convolutional neural networks (CNN).

The performance of any CNN greatly depends on the quality of the training dataset. Generally when training CNN, we choose subset of the dataset randomly. Random selections may perform well, when no constraint on the quality or number of images in the subset. In this paper, we propose image selection method for the semi-supervised classification and compare it to random selection.

The random selection of subset of dataset ignores the properties of the data and produces a repetitive dataset, which is not useful for classification. Consider an example of classification of cats and dogs. In Fig. 1., all images are of cats. The first two images seem to be visually more similar than the third. Consider two classification models, one is trained on first two images and the second model trained on second and third image. Obviously, the second model would be more robust than first. This factor is not important, if we have no restriction on number of labeled images. However, in case we have limitation, we should consider a selection method. We propose *image selection* method that keeps variety in the dataset. In our method, we train an autoencoder, first to extract the most important information in an image, and second to group images with similar information together.

Namely, preparing labeled images for training is a crucial for training a CNN network for classification. However, because the manual labeling requires manpower, the task can be expensive and laborious. Furthermore, in the case of retail stores, staff and customers wear different clothes according to the season. Sometimes, different dress code of staff between different stores. Therefore, labeling large number of images from different season and different stores is not a good option.

Figure 1: A comparison of different images of cat

Semi supervised learning [2] is an attempt to solve the issue of a large number of labeled data. It uses a combination of a large number of unlabeled data and a small number of labeled data. The labeled data is used to predict an artificial label for unlabeled data. The artificial label along with true label is used to prepare a final model for deployment. Several algorithms have been developed using this approach [1, 3, 5]. Here, as we focus more on the data selection part, in our experiments, we use a more simpler algorithm, called *pseudo labeling* [5]. Apart from recent progress in semi supervised learning, other research investigated the impact of the dataset on the algorithm [8, 7, 9].

The rest of the paper is organized as follows. In Sect. 2.,

Contact: Tushar Singh, AWL Inc., Hokudai Business Spring 105 Kita 21-jo Nishi 12-2 Kita-ku Sapporo 001-0021, 070-4377-3096, tushar@awl.co.jp

we demonstrate our method, namely the selection based on autoencoder and clustering. In Sect. 3., we explain our experiments on the retail dataset. We discuss our results in Sect. 4.. In Sect. 5., we conclude with remarks on future directions of research.

## 2. Method

Our proposed method broadly consists of the following three components.

### 2.1 Autoencoder convolution neural network

Autoencoders are a type of deep neural networks, which has wide applications in computer vision. For the purpose of image selection, we are more focused on the data compressing and data coding properties of autoencoders.

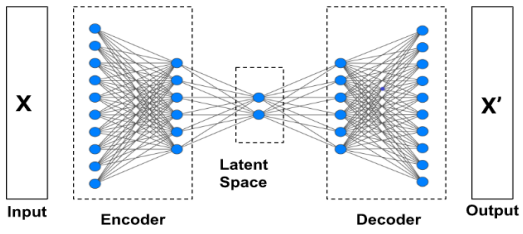


Figure 2: Representation of Autoencoder Network

The basic autoencoder consists of two sections: encoder and decoder, which are connected via a bottleneck layer as depicted in Fig. 2.1. Let $X$ be a vector of an input image. The encoder compresses $X$ into a vector $\mu$ of lower dimension. Vector $\mu$ is called a latent representation of $X$ in a lower dimension space, called latent space. The decoder recreates an image vector $X'$ from the latent representation $\mu$. Ideally, the input $X$ and the output $X'$ of the network should be the same. The latent space $\mu$ is a compressed low dimensional representation of the input $X$.

The autoencoder convolutional neural network is trained such that the output image vector $X'$ is as similar as possible to the input image vector $X$. Namely, the training minimizes the loss function $L = ||X - X'||^2$. A low value of $L$ ensures a meaningful latent representation of $X$.

In this research, we train the autoencoder convolutional neural network to obtain a latent representation that depicts the features that are useful for the classification. In recent year, new variations of autoencoders have been introduced for different applications [6, 10]. The basic type explained above is used in our experiments.

### 2.2 K-means Clustering

Clustering [4] is a classical method to establish relationships among various data in a vector space, for instance relationship between image vectors. K-means is commonly used clustering algorithm. This algorithm tries to partition the images vector space into K subsets. From autoencoder, we get $\mu$ for each image. We then apply K-means clustering algorithm on $\mu_1, \mu_2, ..., \mu_N$, where $N$ is the total number of unlabelled images. Here $K$ is dependent on the number of training images required.

### 2.3 Image selection

This is the last part of our selector algorithm. The output of the previous sections are K clusters, containing similar images. From each cluster, we pick one image. It ensures diverse images in the initial training set. The selection of an image from the clusters can be performed in two ways. In the first way, any random image is picked from a cluster. In this way sometimes boundary images have chances of overlapping. To further ensure diversity in the labeling sample, we pick the image nearest to the centroid of each cluster. In the scope of this paper, we show experiments that follow the first way.

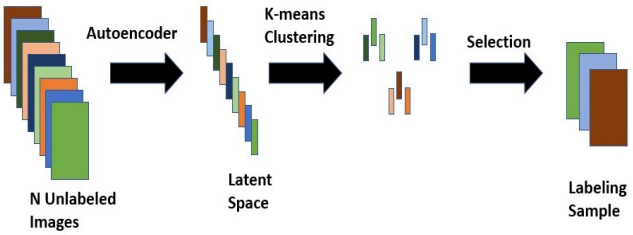


Figure 3: Overall Methodology

## 3. Experiments

This research was conducted for the drug store chain Satudora. The goal is to classify staff/customer of Satudora. Our method has two main step. In the first step, we apply the *selector* to select images with meaningful properties that we call representative images. In the second step, we proceed for the semi-supervised binary classification which is based on the pseudo-labeling technique.

### 3.1 Selector

#### 3.1.1 Dataset

Since our research was conducted for the drug store chain Satudora, We prepared Satudora staff/customer dataset. Video streaming from security cameras were recorded. Then, video is splitted into image frames, on which a state of art person detection model is applied to extract staff and

(a) Original images

(b) Reconstructed images

Figure 4: Original images (1st row) and images reconstructed by the decoder (2nd row)

customer images, or more precisely bounding boxes. The dataset is collected on a daily basis. We use 10,000 unlabeled images from one day for training the selector.

#### 3.1.2 Implementation

We perform image selection based on the autoencoder. We train the auto-encoder network to generate the latent representation of the unlabeled images. For all autoencoder based experiments, we use 2D convolutional, max pooling, batch normalization and upsampling layers to build a network. Adadelta optimizer [11] is used with initial learning rate and decay factor 1 and 0.95, respectively. We also adopted the mean square error as loss function. The input images for training are reshaped into $(64 \times 64 \times 3)$ and normalized.

After completion of autoencoder training as explained above, we obtain images in the latent space of dimension 512. In other words, the input images are transformed into a compressed representation. One of the strong points of the autoencoder is the ability to reconstruct the images from the latent representation. And this is possible thanks to the decoder part (see Fig. 2.1). Consequently, among different latent representation, we are able to select the best one for our experiment. Figure 4(a) shows a sample of original input images, and Fig. 4(b) depicts the reconstruction of these images from the latent representation.

Next, we group the compressed images into clusters using K-means clustering algorithm. In such a way, similar images are grouped in one cluster. To decide similarity between images, we use the Euclidean function.

### 3.2 Classification

#### 3.2.1 Dataset

From each cluster generated by the selector, we randomly pick up one representative image. For our experiment, we start with a small set of 125 representative images for training, where the number of staff and customer images are almost balanced. In the next step, we select larger sets of images of 200, 500 and 1000 as shown in Tab. 1. We manually give labels to these images for training, i.e. staff or customer. Validation set consists of 1787 images, which were prepared based on timestamp and kept completely unseen for the model. Specifically, we collect training data on day $D$. Then next day's data $D+1$ are used for validation.

#### 3.2.2 Implementation

For the training of semi-supervised classification, we utilize the MobilenetV2 network architecture. We also use a pre-trained model imagenet weights for initialization. Data augmentation like horizontal flips, rotation, vertical and horizontal shift employed during training. Adam optimizer and binary cross entropy loss function are considered for the training. To deal with small dataset and variance error associated with model, pseudo-labelling and k-fold learning techniques are examined. Based on experiments, we have decided the batch size of 128 along with 3 folds learning. We have added fixed size of 1000 images for pseudo-labeling from unlabelled set for all experimentation results. For each label, i.e. customer or staff, we take the top 500 strongest predictions for training the classification after pseudo-labeling. The model is trained for 25 epochs in all three folds before pseudo labeling and evaluated on validation set. Further, model is trained on combination of initial training images plus 1,000 pseudo labeled images for 50 epochs.

Table 1: F-1 score for classification model on validation set

| Selector | Training | 125 images | 200 images | 500 images | 1000 images |
|---|---|---|---|---|---|
| Random selection | Before Pseudo-Labeling | 0.82 | 0.66 | 0.84 | 0.82 |
| | After Pseudo-Labeling | 0.83 | 0.83 | 0.88 | 0.92 |
| Autoencoder | Before Pseudo-Labeling | 0.85 | 0.85 | 0.89 | 0.90 |
| | After Pseudo-Labeling | 0.90 | 0.88 | 0.89 | 0.92 |

Note: The pseudo-labeling adds 1,000 images.

## 4. Evaluation and results

We have evaluated our classification results on random selection and autoencoder selection method. Training set consists of different numbers of training images selected by selector method for training semi-supervised binary classification model. The results are reported for staff/customer classification on validation set in Tab. 1. The F-1 score metric was considered as our evaluation metric for the classification task. Experiments are conducted on four different size of label data to check selector performance based on classification results.

Throughout our experiments, our proposed method achieved higher performance on validation set for all size of train dataset comparing to random selection method. Significant margin in F-1 score can be viewed on smallest size of training data (125 images) after pseudo labeling which is approximately 7% higher than random selection method. Meanwhile, with the same architecture, the proposed method performed well even when we increase the size of training dataset for classification model. The results before pseudo labeling in autoencoder selection method indicates the effectiveness of selector algorithm with varying size of training dataset. As train dataset size increases, clusters formed by selector get increase which has direct impact on classification score before pseudo labeling. We have achieved 90% score before pseudo labeling for 1000 images of training set. Since training set and validation set were separated based on timestamp, we avoided the over-fitting problem.

## 5. Conclusion

We presented a method for selection of images based on the autoencoder. The obtained images are representative images that depict variety in properties. We performed semi-supervised classification on the selected images. The results show that classification of selected images using autoencoder has higher accuracy than random selection. As future direction, we plan to perform further experiments with other types of auto-encoders such as variational autoencoders, adversarial autoencoders. Furthermore, we

plan to generalize our method to other types of AI models, such as multi-classification or face estimation.

## 6. Acknowledgement

We sincerely thank Kaname Ujiie for his unfailing assistance and support.

## References

[1] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 5050–5060, 2019.

[2] Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009.

[3] Emily Denton, Sam Gross, and Rob Fergus. Semi-supervised learning with context-conditional generative adversarial networks. *arXiv preprint arXiv:1611.06430*, 2016.

[4] AM Fahim, AM Salem, F Af Torkey, and MA Ramadan. An efficient enhanced k-means clustering algorithm. *Journal of Zhejiang University-Science A*, 7(10):1626–1633, 2006.

[5] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 2, 2013.

[6] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

[7] Mohammad Peikari, Sherine Salama, Sharon Nofech-Mozes, and Anne L Martel. A cluster-then-label semi-supervised learning approach for pathology image classification. *Scientific reports*, 8(1):1–13, 2018.

[8] Mohammad Peikari, Judit Zubovits, Gina Clarke, and Anne L Martel. Clustering analysis for semi-supervised learning improves classification performance of digital pathology. In *International Workshop on Machine Learning in Medical Imaging*, pages 263–270. Springer, 2015.

[9] Eftychios Protopapadakis, Athanasios Voulodimos, and Anastasios Doulamis. On the impact of labeled sample selection in semisupervised learning for complex visual recognition tasks. *Complexity*, 2018, 2018.

[10] Yunchen Pu, Zhe Gan, Ricardo Henao, Xin Yuan, Chunyuan Li, Andrew Stevens, and Lawrence Carin. Variational autoencoder for deep learning of images, labels and captions. In *Advances in neural information processing systems*, pages 2352–2360, 2016.

[11] Matthew D. Zeiler. ADADELTA: an adaptive learning rate method. *CoRR*, abs/1212.5701, 2012.