VIETNAM NATIONAL UNIVERSITY, HANOI
UNIVERSITY OF ENGINEERING AND TECHNOLOGY

Pham Truong Giang

Student Thesis

LEARNING EMBEDDINGS FOR RECOGNIZING HAN-NOM CHARACTERS IN VIETNAMESE HISTORICAL BOOK

HA NOI - 2022

VIETNAM NATIONAL UNIVERSITY, HANOI
UNIVERSITY OF ENGINEERING AND TECHNOLOGY

Pham Truong Giang

Student Thesis

Learning Embeddings for Recognizing Han-Nom Characters in Vietnamese Historical Books

Supervisor: Dr. Ta Viet Cuong,
MsC. Kieu Hai Dang

HA NOI - 2022

AUTHORSHIP

I certify that this work is original and was written wholly by me under the supervision of Dr. Ta Viet Cuong and MsC. Kieu Hai Dang, and that it has never been submitted for a degree. All required reference materials and related works are properly cited, and the author of this thesis ensure that no plagiarism is present.

Hanoi, 05-05-2022

Author

Pham Truong Giang

ABSTRACT

The historical Han-Nom books which were written during pre-modern era of Vietnamese contains are important sources of cultural and historical values. One of the main challenge of analysis task of Han-Nom books is to detect and recognize the character. There are several open sources which can be used for working with Han-Nom books. However, these tools are usually designed to work with Chinese character rather than Han-Nom character which makes them not suitable for analyzing the historical Vietnamese Han-Nom books. Moreover, because of the evolving and mixing of language, the historical Han-Nom characters change overtime and it is difficult to design a universal detector which can work with a wide range of books.

In order to resolve the above issues, our works employ a few shot learning approach which could be trained based on a small amount available data. We propose a way to recognize the detected characters given only a few samples. For dealing with the small amount of available data, we employ an representation learning approach. Instead of training a classifier, our network learn an efficiency embedding which can be used to retrieve the candidate characters from the list of labeled characters.

We evaluate our method on a wide range of historical Han-Nom books which include the printing-style, Han character, handwriting-style character and Nom character. For classifying characters, our learning represenation approach out-performs the traditional classifier training on the domain of few-shot learning. Our proposed method could improve the top 1 and top 5 accuracy substantially in comparison to train a standard classifier. Moreover, we build a small web-based demo which can be used for detecting and classifying Han-Nom characters in books automatically.

Contents

List of figures

List of tables

CHAPTER 1. INTRODUCTION

1.1 Motivation

Sino-Vietnamese characters (Vietnamese: Hán Nôm) are Chinese-style characters perused as either Vietnamese or as Sino-Vietnamese. When they are utilized to compose Vietnamese, they are called Nôm. The same characters may utilized to type in Chinese.

Han-Nom comprises of two writing systems: Hán scripts and Nôm scripts, where Hán script is the Chinese characters, Nôm script is the (basically) Chinese characters used to represent Sino-Vietnamese vocabulary and some native Vietnamese words. The meaning of the name "Nôm" is that this is the language used to record the voice of the Southern people (Vietnamese people), as opposed to the Chinese characters used by Northern people, ie Chinese people. As a result, Nôm script inherits many existing words from Chinese scripts and other words represented by new characters created using a variety of methods.

Han Nom characters are a cultural heritage of Vietnam, it plays a particularly important role in literature throughout the history. Institute for the Study of Han Nom Vietnam is currently storing hundreds of valuable Han-Nom documents in the study of the ancient Vietnamese in many fields: literature, ideas, philosophy, art, language, law, history, morality, etc. However, there are only a few people who can read and write Han-Nom nowadays, making the research of Sino difficult. Therefore, applying technology to automatically recognize and analyze Han-Nom characters is of great importance and has wide practical impacts on the research of ancient Vietnam. Some of the book samples are shown in Figure 1.1. The six books include printing-style Han, writing-style Han and Nom characters.

1.2 Approach

There are a number of challenges in building a system that can detect Han Nom characters. Although many Han Nom documents are collected, most of the documents are currently unlabelled, and most of supervised machine learning model requires large amount of labeled data in order to operate efficiently. Moreover, Han Nom characters has been continuously updated throughout the history so when working with a new document, there is a good probability that we encounter one or more new characters.

Figure 1.1: Han Nom books samples

For all of the above reasons, our approach focuses on building a system being able to observe a small amount of Han Nom characters but can already generalize well to unseen data. To reach our purpose, representation learning is used, which is a method to learn a feature vector of fixed size from each input image. The feature vector is then compared together using different metrics to determine the class of the input image.

The contributions of our work include:

- Employing two representation learning techniques for recognizing Han-Nom character: Triplet and Manifold Mix-up to extract features' vector from Han Nom characters' images. We then adopt K Nearest Neighbor (KNN) and Power Transform - Maximum A Posteriori to match features vectors and classify Han Nom characters.
- Building a pipeline to aggregate embedddings information of Chinese characters' font images into the Han Nom book dataset so that the model can generalize better on new dataset without much more training effort.
- Building a few-shot dataset with character from old Han-Nom book for validating our approaches.

1.3 Outlines

Our report is structured as followed:

- Chapter 1: Introduction. This chapter presents motivation for applying technology in recognize characters in Han Nom documents and shows the importance of applying few-shot in document analysis.
- Chapter 2: Related works. This chapter briefly describes the related work.

- Chapter 3: Methods. This chapter describes in detail our method for text recognition in few-shot settings.
- Chapter 4: Experiments and Results. This chapter shows training details, illustrates the result and discusses it.
- Chapter 5: Models Deployment. This chapter presents technology and results on deploying the models which is built upon the results of our experiments.
- Chapter 6: Conclusion and future works.

CHAPTER 2. RELATED WORKS

As mentioned in the Introduction chapter, our work involves Han Nom characters classification, in which we acquire the use of some few-shot technique to classify the characters' images that has been localized. Chapter 2 presents previous works on Han Nom characters recognition and background knowledge on Deep Network, Few-shot learning.

2.1 Overview on Han Nom characters classification

This section review some of the previous works done in the field of Han Nom characters recognition.

In [24], the authors present a method to collect hand-written Han Nom characters from books and classify them. The classification method is described in the following steps:

- **Step 1**: Produce a hand-crafted feature vector of size 160 representing each image. The feature vectors are created using the gradients vector calculation [12] from input images.
- **Step 2**: Use an existing Chinese OCR to classify the characters. However, they do not accept all of the OCR classifications because as stated in the paper, only around half of each Nom document is in Chinese. As a result, they only select out only 100 classes from all classes based on the distance of the embeddings calculated in step 1 to the prototypes' embeddings of the characters.
- **Step 3**: From the top 100 classes, a method called MQDF2 [8] is used to calculate the score of the character images in relation to the top classes. If the score is low enough, a character image will accept the label. If not, the label is rejected. (step 3 is visualized in figure 2.1)
- **Step 4**: For the rejected images, K mean clustering is executed on their embedding space, then a human labeler will check the validity of the clustering and input the cluster label by hand.

Different papers such as [22], [23] provides similar ways of digitalizing Han Nom documents. Although the methods in these papers achieve good results, some drawbacks should be mentioned. First, the feature vectors are obtained by a fixed methodology, thus

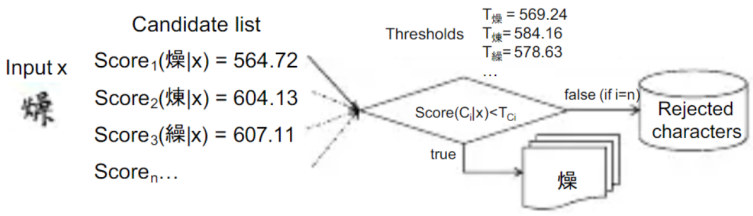


Figure 2.1: Labelling and Rejecting characters [24]

it is not flexible and might not be as good as feature vectors output by a deep learning model, as going to be explained in the next sections. Second, using a Chinese OCR makes the model only able to predict Chinese characters and unable to update new characters to the classifier easily. Third, rejected images are labeled by hand, taking into account that Nom characters' alphabet contains 30k plus characters (many of them are not in the Chinese alphabet), the labeling process will be challenging.

There are other papers which use deep learning models in categorizing Han Nom characters. In [15], the authors try to classify over 30k plus characters using traditional deep learning model, which output the classes probability. This is a good approach since the number of Nom labels are big, the authors also cover the case of class imbalance. However, in cases when there is a new label, the model will have to be retrained completely.

2.2 Deep Network

This section will introduce to the concepts of Deep Neural Network, different components of a Deep Neural Network, a variation of Neural Network called Convolutional Neural Network (CNN), and explain the structure of the deep models we have used in the project.

2.2.1 Deep Neural Network

Deep Neural Network [4] is often seen in supervise learning scenarios. In supervise learning, we have datasets with (x, y) input, output pairs. The goal of machine learning

is to find a function f, so that for each input x, $f(x) = y$. Since finding the exact function f is often impossible in reality, all machine learning models try to find another function $f^* \sim f$ so for each input x, we have $y \sim f^*(x)$. In fact, each function f^* not only takes input x as the input to the function, deep learning models often also have their own set of parameters θ, when the correct set of θ is applied, we can find our optimal function f^*. And the goal of most supervise learning models as well as deep learning models is to figure out the best θ through different model architectures and techniques. Overall, we can formulate the problem of supervise learning for deep learning as given input, output pairs (x, y), find a function f^* and a set of parameters θ so that $y \sim f^*(x; \theta)$.

The reason why it is called Network is because in Neural Network, function f^* is a compilation of different functions $f^*(x) = f^n(f^{n-1})(f^{n-2}(...f^1(x)))$. Each function $f^k; k \in 1, 2, 3, ..., n$ is called a hidden layer. A function f^k is normally a linear transformation combined with an activation function. An activation function are used to increase the deep learning model non-linearity, since (x, y) pairs in every dataset do not always have linear relationship. General structure of a simple Neural Network is found in figure 2.2 while different activation functions are shown in figure 2.3

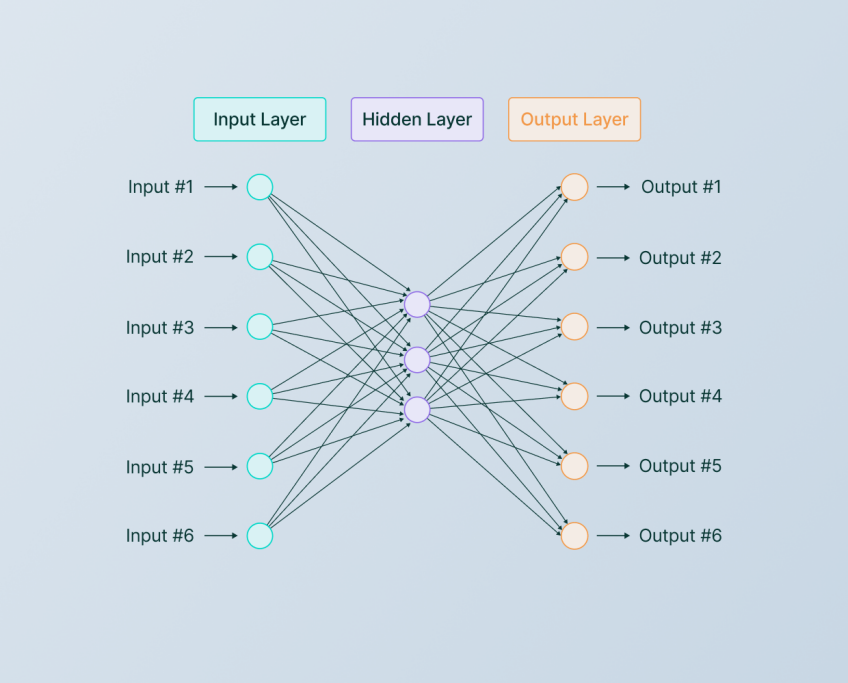


Figure 2.2: Simple Neural Network Architecture, taken from [1]

As explained above, the task of a typical deep neural network is to find the set of θ so that $y \sim f^*(x; \theta)$. In order to measure the similarity between the output and our prediction function, a Loss function is used. The loss function is defined as $L = Loss(f(x), y)$. The smaller the Loss, the better our learned function is. The method to step by step modify the set of θ so as to minimize the Loss function is called Gradient Descent. For example, assuming that we have of θ_0, which produce loss L, then Gradient Descent will update θ_0 to θ_1 according to formula 2.1. The formula will run iteratively in the hope that after some iterations, θ will converge and loss L will be at its local or

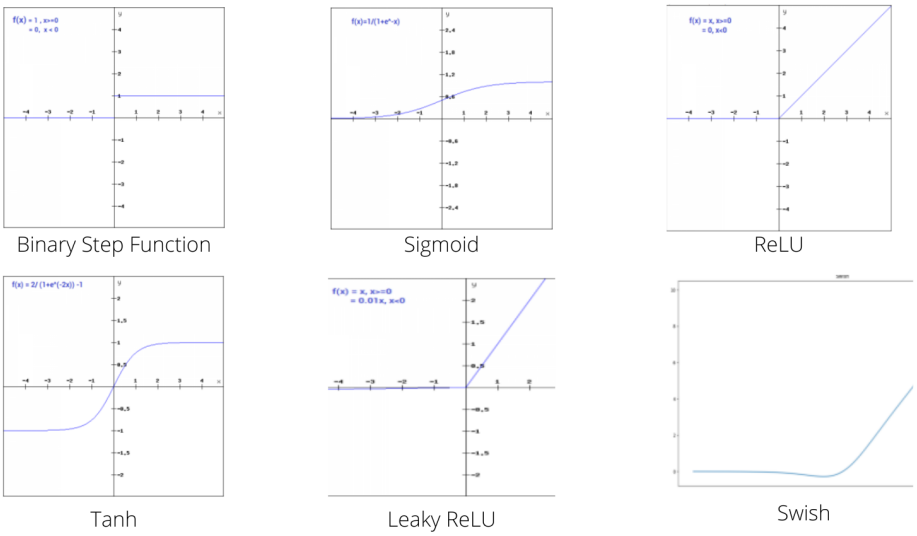


Figure 2.3: Different activation functions [20]

global minimum. Of course, in order to apply the formula 2.1, the loss function should be differentiable. And the process of updating θ is called the training process.

$$\theta_1 = \theta_0 - \frac{\nabla L}{\theta_0} \tag{2.1}$$

In traditional machine learning, we often divide the dataset into training set and test set (sometimes we also have the validation set). The training set will be used for the process of updating the θ parameters. The test set is used to validate the function learned on the training set. Then when training and testing the model, a few patterns will emerge. In the first scenario, the model achieves high result on both training set and test set, then we might have trained a good model since it can generalize well even on unseen data in the test set. In the second case, our model gets low results on both training set and test set, which is called under-fitting, it happens when the model cannot adapt to both the training set and test set. The third scenario is the model gets high result on the training set but low result on the test set, we call the pattern to be over-fitting. In case there is too little labelled data in the training set, we can fit the training set well, however the training set is not enough to represent the distribution of the dataset, thus get low result on the test set. This over-fitting case mentioned above happens to be the problem that few-shot learning wants to tackle, the subject of which I will discuss deeper in section 2.2. Under-fitting, Over-fitting and Appropriate fitting are visualized in classification is visualized in figure 2.4.

However, there are some problems when applying simple neural network model as described in figure 2.2 to the image dataset since the input to simple neural network architecture is a flat vector. However, in image dataset, the area surrounding a pixel tells

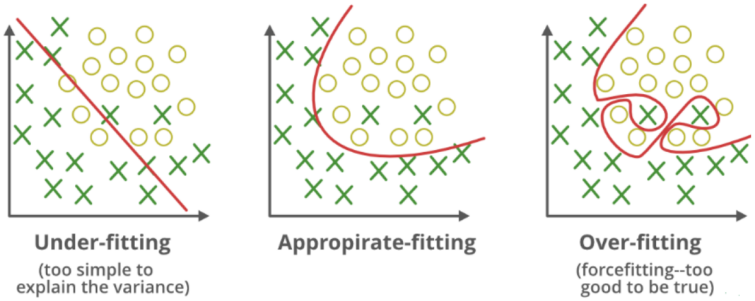


Figure 2.4: Different model fitting scenario [2]

something about that pixel as well, so flattening the image will cause losses to the images' information. So for the image dataset, Convolutional Neural Network architectures are leveraged.

2.2.2 Convolutional Neural Network

Convolutional Neural Network [16] is very similar to traditional deep neural network architecture except they do not flatten the input immediately, but instead make use of the convolution operation to extract information not from individual pixel (in the case of image dataset) but from their neighborhood as well. The following passages will clarify the concept of convolution operation and pooling operation, two typical operations in Convolutional Neural Network. I will also present about the basic architecture of Residual Neural Network, the backbone I use in all of my models.

Convolutional Operation

Convolutional Operation involves the use of a small kernel (normally of size $(2k + 1) \times (2k+1)$). The kernel slides across the input images or feature maps to output another feature map as shown in figure 2.5. The output feature map then goes through activation function to increase the non-linearity of the model as explained in section 2.1.1.

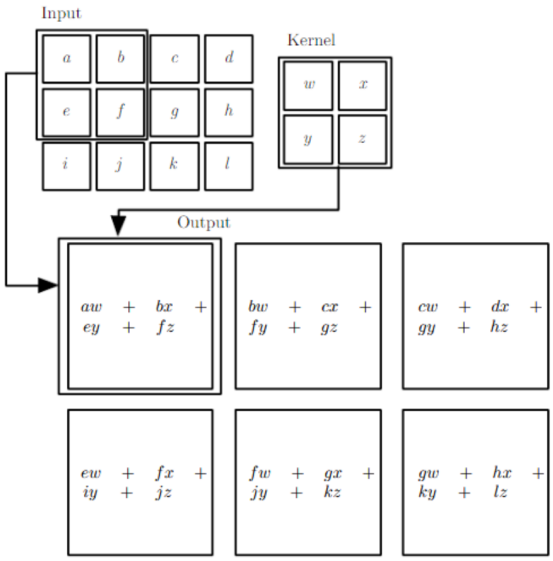


Figure 2.5: Different model fitting scenario [4]

Pooling Operation

In a typical Convolutional Neural Network Architecture, the input goes through several convolutional block with activation function to extract different features of the image. Then the pooling layer is applied. The output feature maps after pooling can be said to contain statistical properties of the previous layer. Pooling also involves a kernel moving through the image, but different from convolutional operation, the output feature map of pooling only depends on the value of the input of pooling layer. Some of Pooling techniques often used are Max Pooling or Average Pooling.

Residual Neural Network

Residual Neural Network architecture is first introduced in [5] after the author realizes the limitation of traditional convolutional neural network. One of them is that after training for some time for model with many layers, gradient explode/vanish will appear, harming the training process. So, the authors add some changes to the architecture. Normally, a layer in traditional convolutional neural network only uses the layer right before it, thus if the layer before it appears to have problem, the latter layers will be affected as well. As a result, a slightly modified structure is introduced as described in figure 2.6. A feature map in layer n will be added to the feature map in layer $n + k$. The intuition is that if feature map in layer n does not have any problem but layer $n + k$ appears to have gradient explode or vanish. The next layer can still be fine since it also uses the layer n. From this basic building block, different architecture was born and used effectively in different settings. My experiment leverages Resnet 50 and Wide Residual Neural Network [31] as backbone to classify between different characters in the Han Nom dataset.

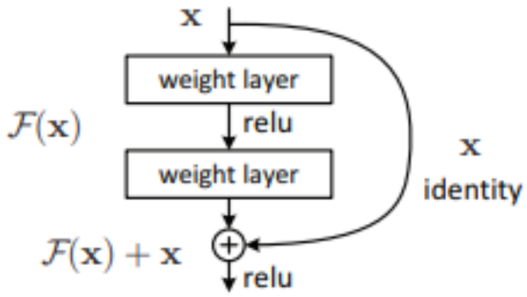


Figure 2.6: Residual Block [5]

2.3 Few-shot learning

The fact that we want to generalize our model to characters in Han Nom books after training with a few examples makes our problem become a few-shot problem. We as humans can learn to recognize a character, whether normal text or Han Nom text only through a few samples. However, most modern deep learning models can only generalize well when given a good amount of data to learn. To bridge that gap between humans and machines, the idea of few-shot learning has emerged. Formally, Few-shot learning is defined as a type of machine learning problem. The Machine Learning model is defined as a model to improve accuracy P on classification task T, given the experience E. The only difference between few-shot learning and other machine learning problem is that in few-shot learning, E has only a few samples [29]. In a normal set-up of few-shot learning, there is a base class (or base dataset) where abundance of data is collected so that the model can learn meaningful features. Apart from the base dataset, there also exists the novel dataset, which contains completely new labels (although not necessary). The novel dataset is also splitted further into support set and query set. Support set contains k labelled data points for each of n classes where k is small and normally 1,5, or 20, we call this kind of dataset n ways k shots. Query set contains data of the same classes as support set and is used for testing the model.

Different ideas have emerged to tackle the problem such as augmenting the data, using different model architectures, or changing the algorithm used with a model. In this part, I will briefly introduce the background on the strategies I have used for the project as well as the previous works done on Han Nom documents analysis.

2.3.1 Transfer learning

Traditional machine learning techniques work under the assumption that the distribution as well as the feature space of the training and test set are the same. When meeting a distribution shift or feature space shift, it is likely that a new dataset with large amount of data should be re-assembled. In the Han Nom project's however, there are only a few training data for each characters and it is not possible to sample more data of the same class with the same image's characteristic. Therefore, I make use of another artificial dataset to pre-train the model as will be explained in later sections. The problem with this approach is that I have to find ways to adapt the model pre-trained on the artificial data to the book dataset, which is called Transfer Learning.

Transfer learning is a technique where a model pre-trained on one task is applied to train another task. In situations where there is a lack of data, the model is pre-trained on another dataset with similar characteristics, so that when applied to the current task, the model can achieve decent results with less data and training.

Machine Learning takes longer and more resources for the model to converge since it only trains with 1 dataset only. On the other hand, the pre-trained model in transfer learning already learns some specific features when training with the previous dataset. As a result, it converges faster with less data. However, the dataset for pre-train (called the source dataset) and the current dataset (called the target dataset) should have some common characteristics. For example, in my experiment, I generated a source dataset of Chinese characters image and with image size of 64×64, closely similar to the book dataset.

Different types of Transfer learning are classified based on 3 main criteria: whether the source and target data is labelled or unlabelled, in the same or different domains, perform the same task or different tasks. Two datasets have the same domain if they have the same feature space or marginal distribution, for example our artificial dataset and book dataset both have the image size of 64x64. And two datasets are used for the same task when they have the same label space, even if their domains are different. Based on the aforementioned criteria, according to [17], 3 categories of transfer learning are listed as below:

- Inductive Transfer learning refers to cases when the target tasks and the source tasks are different but the domains are the same. Inductive Transfer Learning can also be classified further. If the source and target tasks both contain features and labels, then it is similar to multi-task learning, else if the source task does not contain label, it is similar to self-taught learning.
- Transductive Transfer learning is quite the opposite of Inductive Transfer learning. While inductive transfer learning datasets have the same domain but different task, transductive transfer learning datasets have the same task but different domain (different in feature space or feature distribution). One more characteristic of Transductive Transfer Learning is that the source dataset contains abundant of data. On the opposite, the target dataset does not contain data with label.

- Unsupervised transfer learning refers to cases when both the source and dataset tasks do not contain labels.

Our approach belongs to the class of Inductive Transfer Learning since our source dataset and target dataset have the same domain but different set of classes. There are different solutions to tackle Inductive Transfer Learning problem.

A solution called Instance Transfer uses the source dataset along with the target dataset in order to boost the performance of the model. Example of Instance Transfer is Adaboost [30].

A solution called Parameters Transfer uses the same settings as Multi-task learning, in which different tasks and dataset share some parameters. Typical examples of Parameters Transfer can be found in [9].

Another solution, which is our approach, is feature representations transfer. From the source dataset, I apply models with different kinds of losses to learn the representations. The models are then fine-tuned to produce best performance on the limited book dataset.

2.3.2 Learning Image Representations

One of the most popular and effective techniques in transfer learning is learning image representations. Traditional deep network contains hidden layers and output layer. The last layer serves the predictive role and the hidden layers learn different features of the input. Learning Image Representation is only different from traditional learning method that instead of using the output layer, it uses different techniques the match the feature vector or feature map of the previous layers. However, of course, we would want our representation to have certain different characteristics. Different characteristics of representation vector requires different types of training. The ways we apply different image representations technique on Han Nom dataset is further explained in chapter 3.

CHAPTER 3. Methods

For general task of character recognition, a reasonable approach is to do a standard supervised learning for classification problem, where each class corresponds to a distinct type of characters. However, the exact total number of characters in Chinese, the most spoken language, is unknown, but they can generally be around 50000 or even more (though most of them are obsolete and rarely used), and as Han-Nom has a deep link to Chinese, we can take it as a rough estimate of the upper bound of the number of Characters that the model should be able to detect. As a result, to train such a large classifier requires a lot of labeled data, which is hard to achieve with Nom as there are only few Nom scholars nowadays. On the other hand, with a fixed number of characters, the model needs to be retrained whenever a new type of characters is recorded, as new Nom documents are recovered. For these reasons, we need a better method that is more flexible and can adapt to new labels with limited data and little changes to the model. Therefore, we decided to approach the problem by learning an embedding space of Han-Nom characters. Such space should encode high-level important information of characters like orientation of strokes and thickness of lines to be able to distinguish between different characters and also to generalize well to unknown characters. With that in mind, we use a model with Triplet loss and Resnet 50 backbone, and a model with Manifold Mix-up loss and Wide Resnet backbone to train on our data. Our data consists of two parts, the first part is the data we generate by using Chinese characters combined with a set of fonts, the second part is characters localized from the Han-Nom books shown in the introduction and labeled by EasyOCR. Chapter 3 aims to explain in details how we use models with Triplet and Manifold Mix-up loss to train the embeddings as well as techniques to match embeddings of the same classes. Different methods to learn the embeddings and match the embeddings are synthesized in figure 3.1.

3.1 Learning Embeddings

As explained in related works, learning embeddings is one of the most popular approach in tackling few-shot learning problems and often achieve high results on different benchmark. In representation learning, we might want our representation to capture the most meaningful and rich features of the image, or we might want the representations of similar class to be close to each other in the embedding space. For the first purpose, our model uses Resnet 50 and Wide Resnet backbone pretrained on different popular dataset and apply to further fine-tune on fonts and Han Nom book dataset, so the model already

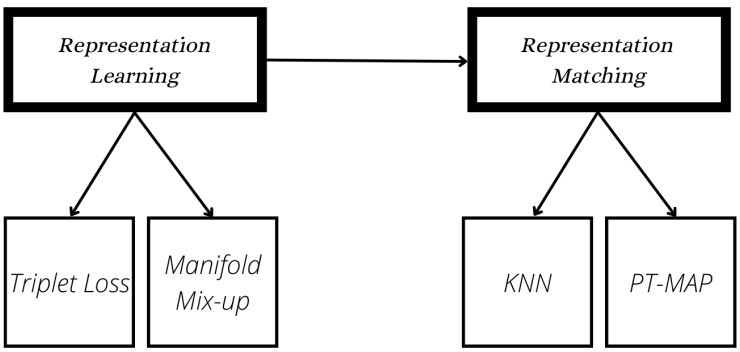


Figure 3.1: Different methods of learning and matching representations

contains rich features in the beginning. For the second purpose, we find and apply two different methods to train on the dataset namely Triplet Loss and Manifold Mix-up. This section will serve as an introduction to the concepts of Triplet Loss and Manifold Mix-up and how we apply these methods on training and validating on the Han Nom documents.

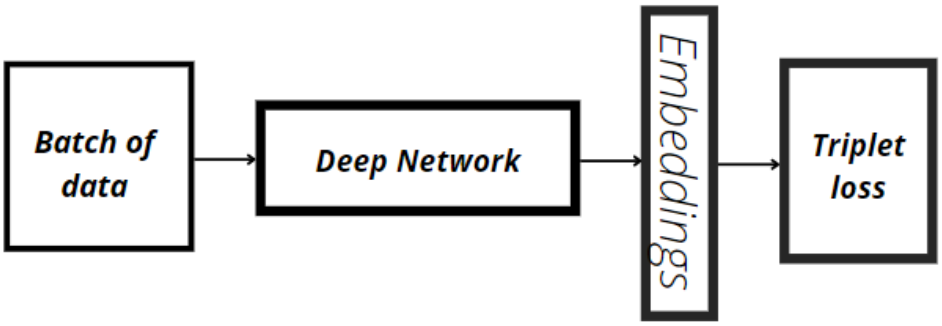


Figure 3.2: Triplet structure

3.1.1 Learning Embeddings with Triplet loss

Triplet learning is the learning method that was first introduced in [19] in the context of face recognition and verification problems, it learns a Euclidean embedding that

abstracts the geometry structure of high level features of the original data. The L2 distances in this learned space directly corresponds to the similarity between original data points. In our problem, similar characters should be close to each other in this embedded space, thus character classification can be done by some clustering algorithms, which in our case is k nearest neighbors.

Triplet method begins with the idea of utilizing an embedding function f, which is commonly implemented as a neural net, to embed an image into feature space, such that the squared distance between characters of the same type, regardless of their writing style and imaging conditions, whereas the distance between pair of different type of characters is large. To make the method an end-to-end learning task, Triplet learning translates this idea into an optimization problem by a special loss function called Triplet loss.

Triplet loss [19] works by choosing a triplet of samples of two similar classes, called anchor and positive, and a sample from another class called negative. Since we want images from the same class to be close in this embedding space and far away to images from different classes, triplet loss directly computes the distance between anchor-positive and anchor-negative pairs and compare the two distances, a margin a is enforced in the loss function to further distinguish between positive and negative pairs. This can be written formally as

$$\|f(x^a) - f(x^p)\|_2^2 + a < \|f(x^a) - f(x^n)\|_2^2$$

The Triplet objective is then the minimum optimization problem of the loss

$$L = \sum_i \max\left(\|f(x_i^a) - f(x_i^p)\|_2^2 + a - \|f(x_i^a) - f(x_i^n)\|_2^2, 0\right)$$

Where x_i^a, x_i^n and x_i^p belong to the set of all possible triplets. This loss function can be easily satisfied by many easy triplets as some negative pairs are easy to tell apart. As a result, these triplets contribute little to the training and can lead to slow convergence and additional mining techniques must be applied to ensure faster convergence. Triplet does this by actively selecting hard positives and negatives, hard positives and hard negatives are pairs that violate the loss function the most, i.e. they make positive distances maximum and negative distances minimum. However, using too many hard examples would lead to slow training as mislabeled and poor condition images dominate the training examples. Therefore triplet uses hard positive pairs and semi-hard negative examples, semi-hard examples are negative pairs that are still further to the anchors than positives but they stay inside the margin. Besides, since positives and negatives are chosen from mini-batches, there should be enough positive examples present in each batch to provide useful anchor positive distances.

3.1.2 Learning Embeddings with Manifold Mix-up loss

In normal deep learning settings, we process inputs and their labels sequentially, however, with the manifold mix up, when training up to a specific layer, the layer and labels will mix up together in a linear fashion. For example, we have input x and x',

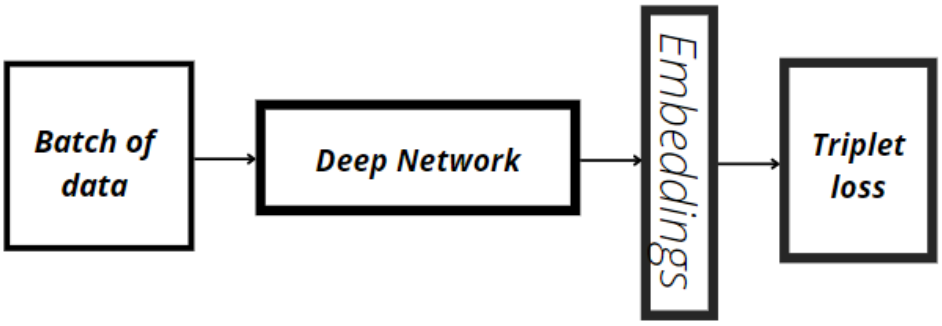


Figure 3.3: Triplet structure

feature vectors at a specific layer are $f(x)$ and $f(x')$, and output y and y'. Then manifold mix-up mixes the feature vectors and their labels together by formula 3.2, 3.3, where y and y' are one-hot vectors. After the mix-up, new batch of data is generated and evaluated.

$$f(X) = \lambda f(x) + (1 - \lambda) f(x') \tag{3.2}$$

$$Y = \lambda y + (1 - \lambda) y' \tag{3.3}$$

After mix-up the data, we use Manifold Mix-up loss as defined in formula 3.4, using symbols defined from formula 3.2 and 3.3.

$$L_{mm} = \mathbb{E}_{(x,y)}(f(X), Y) \tag{3.4}$$

Manifold Mix-up technique is first introduced in [25] as an regularizer to tackle contemporary deep learning problems. For example, they realize that in regularizers using weight-decay [13], dropout [27] and batchnorm [7], the decision boundary is often not smooth enough and thus most predictions are made with high confidence, even with samples that are close to the decision boundary, however, intuitively, we want the samples being close to decision boundary should be predicted with less confident. This weakness makes normal deep learning settings with aforementioned regularizers susceptible to adversarial attacks, even a small change which cannot be captured by the human's eyes can cause the model to predict totally different class with high confidence. Manifold mix-up technique, by smoothing the decision boundary, partly restrains the effect of adversarial attacks. The difference in decision boundaries when compared between Manifold Mix-up and other regularization techniques can be visualized in figure 3.4.

There are different reasons that we decide to experiment with Manifold Mix-up in addition to Triplet loss model. First, in [25], the author has plotted out the decision

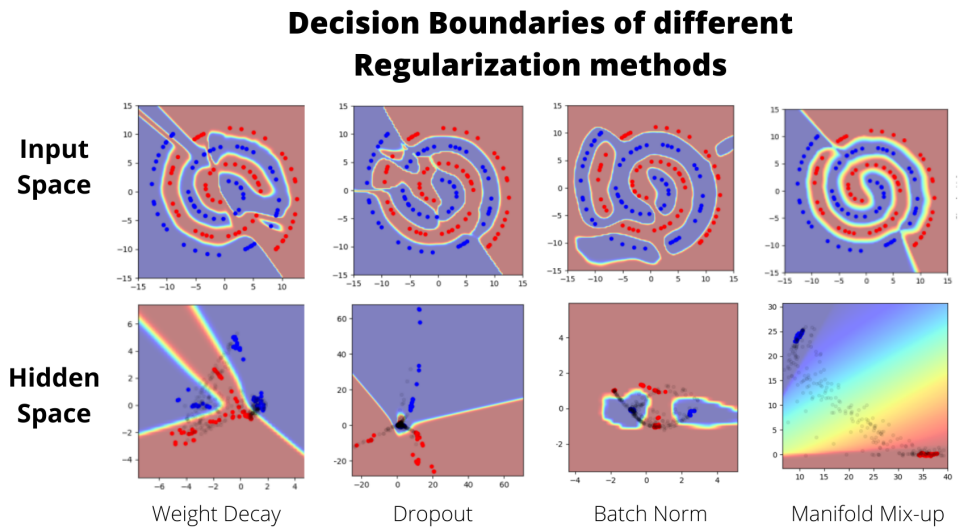


Figure 3.4: Comparison of resulting decision boundaries after using different types of regularization techniques. There are 2 types of decision boundaries visualized here: the first one is from the input space, the second one is from the hidden space (decision boundaries from one layer of the model). The visualizations are taken and edited from [25]

boundary in the hidden space of the model using manifold mix-up as seen in figure 3.4. From observation and proof from the paper, we can conclude that Manifold Mix-up not only improves decision boundaries in the input space, but it also does well on the hidden space and from its representation, we can confidently separate between classes with even simple embeddings matching techniques like KNN, which is partly how Manifold Mix-up is applied in our experiment. Second, it is also stated that model trained with Manifold Mix-up techniques perform well when there is unexpected transformations of the input images, which is suitable since different Han Nom books have different types of page styles and the bounding boxes of the characters are not always nicely cropped.

Because of the ability to learn good representations as explained above, manifold mix-up is also leveraged in few-shot learning settings [14]. Few-shot learning set-up often includes Base Classes and Novel Classes. The Base classes contains sufficient data and labels so the model, or feature extractor can learn meaningful representations. The Novel Classes contain only a few training data with labels belonging to classes that are different from Base Classes, and used to test the representation learned by the model having been trained with Base Classes. Our experiment uses the same settings, we use an artificial dataset from Chinese characters and fonts with abundant data as Base Classes to train Manifold Mix-up on. In the training phase, we train on Base Classes in the same way as training traditional machine learning models, with the last layer as a one-hot vector representing probabilities for each class. In the testing phase, we fine-tuned on the characters' images like before, then removed the last layer and used the penultimate layer as embeddings to test the model's accuracy.

By definition, the mix-up happens after the input has been trained through some

layers (the number of layers can be 0). However, we do not know if mixing-up at which layer gives the best result. So we used the charting the right manifold method proposed in [14], a method that randomly chooses which layer to mix-up the data for each iteration. The structure of Manifold Mix-up experiment is shown in figure 3.5.

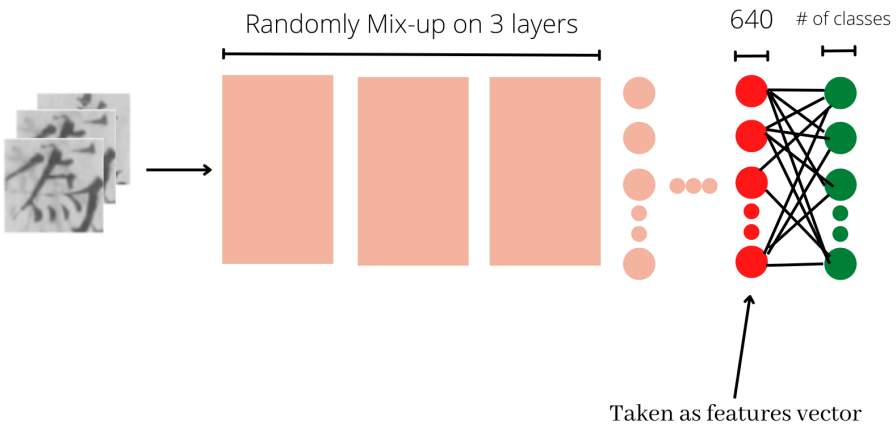


Figure 3.5: Manifold Mix-up model architecture. The first three layers represented by pink rectangles are layers where mix-up between data happen. In the training phase, we use the last layer to classify between classes. In the testing phase, we use the penultimate layer of size 640 as representation vector for the image.

3.2 Matching Embeddings

After getting all the embeddings for the support as well as query images from Han Nom dataset. We need a way to infer the correct class of the query embeddings obtained from the feature extractor. We can look at this problem from multiple perspectives, each perspective gives us different ways to infer the result from embedding space. In one perspective, we consider these embeddings as points in the embedding space, the closest point to the target might be the class the target belongs to. In statistical perspective, each element in the features vector of an image has its own distribution, and if we can have some insights on each element of the features vector, we might be able to infer the correct class, since we assume that each class should have different features vector. We use K Nearest Neighbors (KNN) with L2 distance to implement the first perspective, and utilize PT-MAP to implement the second perspective. This section will be dedicated to explaining the concepts and usage of KNN and PT-MAP in our attempt to classify the Han Nom dataset.

3.2.1 K Nearest Neighbor (KNN)

K Nearest Neighbors [18] is one of the most simple and popular method for classification. With K Nearest Neighbors, we make a prediction by selecting the K closest labelled data points to the target based on some metric. Our experiment leverages L2 loss for K Nearest Neighbors algorithm.

K Nearest Neighbors method is based on an assumption that the feature vectors of objects in the same class are close to each other. This assumption holds true for our representation learning method. In triplet loss model, we try to train the embeddings by keeping embeddings of the same class close together while pushing embeddings of different classes further away. Manifold Mix-up is no difference since as visualized in Manifold Mix-up session, even in its hidden layers, Manifold Mix-up still outputs good decision boundary that clearly separates different classes. KNN decisions on top K Nearest Neighbors by L2 loss are visualized in figure 3.6

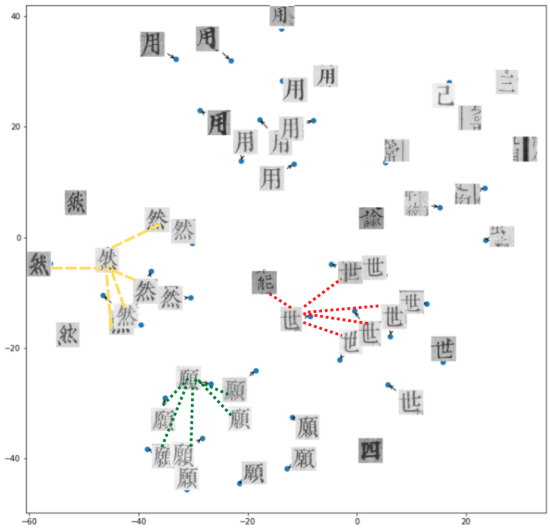


Figure 3.6: Top k nearest neighbors of different characters trained by Manifold Mix-up model. The colored dotted lines are top 5 selections by the specified characters.

3.2.2 PT-MAP

As explained in the previous sections, embeddings training is one of the most appropriate and popular ways to deal with few-shot learning problems. It is also leveraged by many few-shot state-of-the-art models. Many works train the feature extractor on Base Classes to make the model easily adapt to Novel Classes. Since it is not easy to always have qualified Base Dataset, many approaches use pretrained backbone on different dataset like Cifar-FS, Image Net, CUB and apply directly to the Novel Dataset. The problem with adapting a pre-trained backbone to the target task is that the data distribution of different tasks are different, which can cause failure to transfer-based few-shot

learning method. As a result, in [6], the authors propose a way to normalize the features vector distribution called Power Transform (PT), and leverage an effective optimal transport algorithm based on Maximum A Posteriori (MAP) for features vector matching. This section will serve as an introduction to PT-MAP (Power Transform - Maximum A Posteriori) and its effect on our transfer-learning model.

Power Transform (PT)

As mentioned in [6], many works on representation learning assume that the distribution of feature vectors to be Gaussian with little to no proof on it. And in reality, the features extractor's output is often skewd rather than Gaussian-like. For example, in our experiment, after going through Wide Resnet backbone pretrained by Manifold Mix-up, the output feature vectors of size 640 are all left-skewd. Different works have already tackled this distribution problem by some statistical methods [28], [11]. However, the authors of [6] argue that for feature vector of unexpected distribution, these statistical methods can only make the training process worse, and they propose to normalize the features vector to Gaussian-like distribution before using the representations for inference.

Inspired by [6], we use the Tukey Transformation Ladder [21] to "Gaussianize" our left-skewd data. The Tukey Transformation Formula is defined in equation 3.5

$$f(v) = \begin{cases} \frac{(v+\epsilon)^\beta}{||(v+\epsilon)^\beta||_2} & \beta \neq 0, \\ \frac{(v+\epsilon)}{||(v+\epsilon)||_2} & \beta = 0 \end{cases} \tag{3.5}$$

In the 3.5 equation, if we assume that the original vector v is left skewed or right skewed, the Tukey Transformation Formula helps making the distribution more Gaussian-like by the value β. By experiment, we find that setting up $\beta = 0.5$ like in [6] transform our data into Gaussian-like distribution. The distribution of a random element in the features vector before and after Tukey Transformation Ladder are shown in figure 3.7.

I apply Power Transform to the representations learned by Manifold Mix-up of the dataset as a whole. As viewed in figure 3.7, the representations of the whole dataset have become Gaussian-like. And we assume that the feature vectors of each class has Gaussian distribution with their own means and standard deviations. The transformed feature vectors are then used as input to MAP algorithm. The distribution of a random element in features vector of three different classes are shown in figure 3.8

Maximum A Posteriori (MAP)

After applying Power Transform to the representation of the whole dataset, each class's features will be assumed to have Gaussian distribution. Since the features distribution is Gaussian-like and we assume the same for every class, we can assume that there exists a center representation of every class. Finding these center embeddings will be beneficial in classification between different classes. If there is abundant data in each

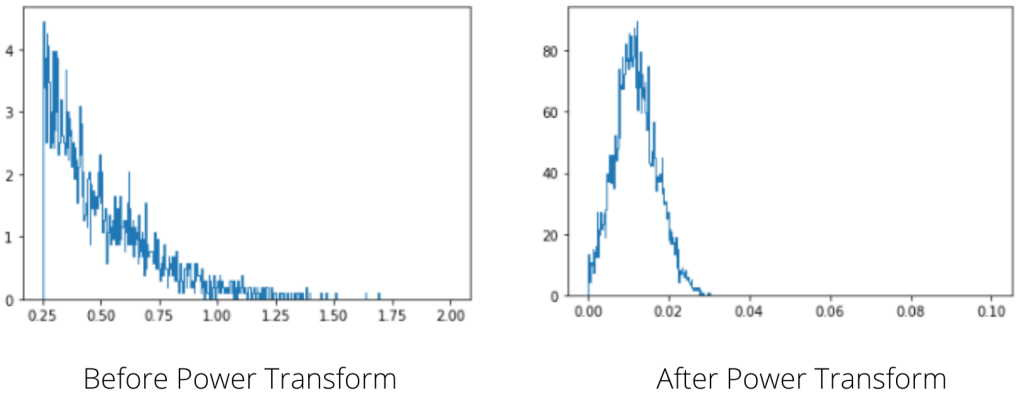


Figure 3.7: Effect of Power Transform on the distribution of a random element from Manifold Mix-up feature vector of size 640 on Han Nom Dataset

class, we can easily calculate the center feature representation by averaging all feature embeddings of that particular class. However, in our situation, we only have 20 labelled embeddings per class (each of the Han Nom dataset's classes have only 20 images with label for training, and 60 unlabelled images for testing), so calculating the center with so few data in one class might not return the center embedding that represent the whole class. The PT-MAP paper proposes an algorithm to iteratively find the center feature vectors of each class, and at the same time updating the query samples' labels so that the center of a class should be as close to the query samples' embeddings as possible.

First, we define a matrix M of size $wq \times w$, where each row sums up to 1 and each column sums up to q. q is the number of query sample in 1 class, w is the number of classes, so wq is the number of all query samples. Each row of matrix M^* can be considered the probability of each image belongs to each class.

We also denote c_j $(1 \leq j \leq w)$ to be the center features vector of class number j; f_i^S $(1 \leq i \leq ws)$ is the feature vector of sample number i in the support set, s is number of samples in a support set class, f_j^Q $(1 \leq j \leq wq)$ is the feature vector of sample number j in the query set, q is the number of samples in the query set class; $l(f_i)$ is the label of feature vector i; L is a matrix of size $wq \times w$, represents the L2 distance of each query feature vector to the center vectors of the classes.

Assume that we have the true center c for every class, we would want the sum of distance between the feature vector of a class to its center to be as small as possible. Formally, the problem is formulated as in formula 3.6

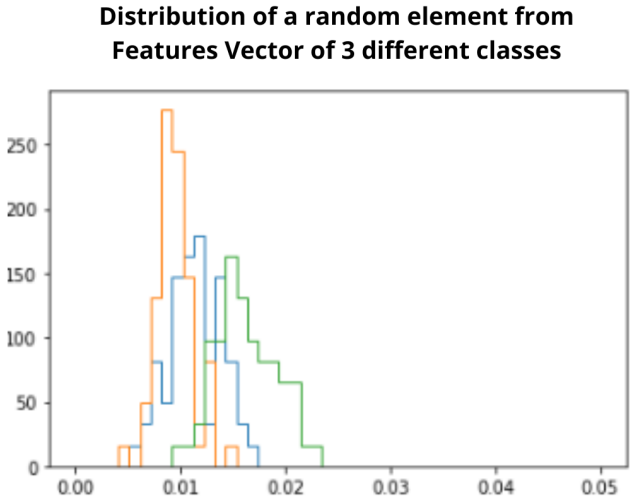


Figure 3.8: Distribution of Random feature of representation vectors from 3 random classes. (from Han Nom book dataset)

$$\begin{array}{lll} \text{minimize} & M^* = \sum_{ij} M_{ij} L_{ij} & \\ \text{subject to} & \sum_j M_{ij} = 1 & i = 1, \ldots, w \\ & \sum_i M_{ij} = q, & j = 1, \ldots, wq \end{array} \tag{3.6}$$

We notice that in formula 3.6, if we divide each element of the matrix to wq, then M become a joint distribution, so we obtain an Optimal Transport problem [26] (Lij is fixed since we are assuming that there are fixed center embeddings for each class). There are several methods to solve the problem. The authors of [6] use Sinkhorn distance [3], an effective distance to calculate the right distribution M^*.

On the other hand, if we already have the right matrix M, we can also re-calculate the center of each class with formula 3.7.

$$u_j = recenter(M, j) = \frac{\sum_{i=1}^{wq} M_{ij} f_i^Q + \sum_{f \in f_S, l(f)=j} f}{\sum_{i=1}^{wq} M_{ij} + s} \tag{3.7}$$

$$c_j = c_j - \lambda(u_j - c_j) \tag{3.8}$$

In formula 3.7, the second line helps c_j does not change too much.

In conclusion, if we have c, we can calculate M, if we have M, we can calculate c. We can initialize the center for class j as $c_j = Average(f)$, knowing that $f \in S, l(f) = j)$. Ideally, by iteratively updating c and M until it converges, we get our final prediction for each query sample by picking the highest probability from each row of M. The MAP algorithm is illustrated in algorithm 1.

Algorithm 1 MAP algorithm

for $iteration = 1, 2, \ldots, w$ **do**
    $c_j = \text{Average}(f), f \in S; l(f) = j$
end for
for $iteration = 1, 2, \ldots, n$ **do**
    $L = ||f - c||$
    $M = Sinkhorn(L, M)$
    $u_j = recenter(M, j)$
    $c_j = c_j - \lambda(u_j - c_j)$
end for

CHAPTER 4. EXPERIMENTS AND RESULTS

Chapter 4 presents our experiments and results. The experiment is divided into multiple steps. First, we train the model on a artificial dataset generated from Chinese characters and fonts so that our model can learn the characters' embeddings better. Second, we sample the most popular characters from all books and split them to support and query set, the support set contains only few images for each characters. On the book dataset, we fine-tune our pre-trained model on the support set, then proceed with making prediction on the query set. The details of the steps are demonstrated below.

4.1 Dataset

In order to learn the embedding for character classification, we use 2 types of dataset. An artificial dataset and a book dataset.

The artificial dataset is created from computer encoded Chinese characters and fonts. The dataset contains around 8k distinct characters associated with 25 fonts. Transformations are then applied to the generated dataset to make the model trained more robust. Some techniques wehave used are:

- **Color Jitter**: a technique that randomly changes the image's brightness, contrast as well as saturation.
- **Random Box Blur**: compared to the original image, the resulting image after box blurred has the pixel values equal to the average of its neighbors' value in the original image.
- **Random Perspective**: randomly changing the image's perspective. Different perspectives of an image are different representation of an object in 2D scenario.
- **Random Affine**: a technique of transformation that preserves planes, lines and points.
- **Random Erasing**: randomly choose a small rectangle in the image and erase it.
- **Random Resize Crop**: randomly crop a portion of the image then resize it back to the size of the original image.

Different fonts for a character are shown in figure 4.1, while the transformation's results for the characters are shown in figure 4.2. Among all the classes, 6000 distinct

classes of characters are chosen to pre-train the model, and 200 novel classes are picked to evaluate the model's representation learning.

Figure 4.1: Different fonts for a character

The book dataset is created from seven historical Han-Nom books, including:

- Phap Hoa De Cuong is a Buddhism book written in the twentieth century by a Vietnamese monk. The images from Phap Hoa De Cuong are captured from a printed book.
- Truy Mon Canh Huan is a Buddhism book, written in Tong Dynasty in China. It contains discipline for the monks to follow. It encourages people to follow a religious life. The images from Truy Mon Canh Huan are captured in a printed Chinese book.
- Khoa Hu Luc is the work of King Tran Thai Tong in the thirteen century. The book can be describe as King Tran Thai Tong's comtemplation on the nothingness. The images are captured from a Chinese printed book.
- Dai Nam Quoc Su Dien Ca is a creation followed the order of King Tu Duc in the era of the Nguyen Dynasty. It is a history book written in Luc Bat, presents the history of Vietnam from King Duong Vuong till the end of the Le Dynasty. The images from Dai Nam Quoc Su Dien Ca are captured from a printed book written in Nom.
- Dai Nam Thuc Luc Tien Bien is published in 1844, and is known as the official records of the Nguyen Dynasty. Dai Nam Thuc Luc Tien Bien is written in classical Chinese. The images from Dai Nam Thuc Luc Tien Bien in the data are captured from a printed book.

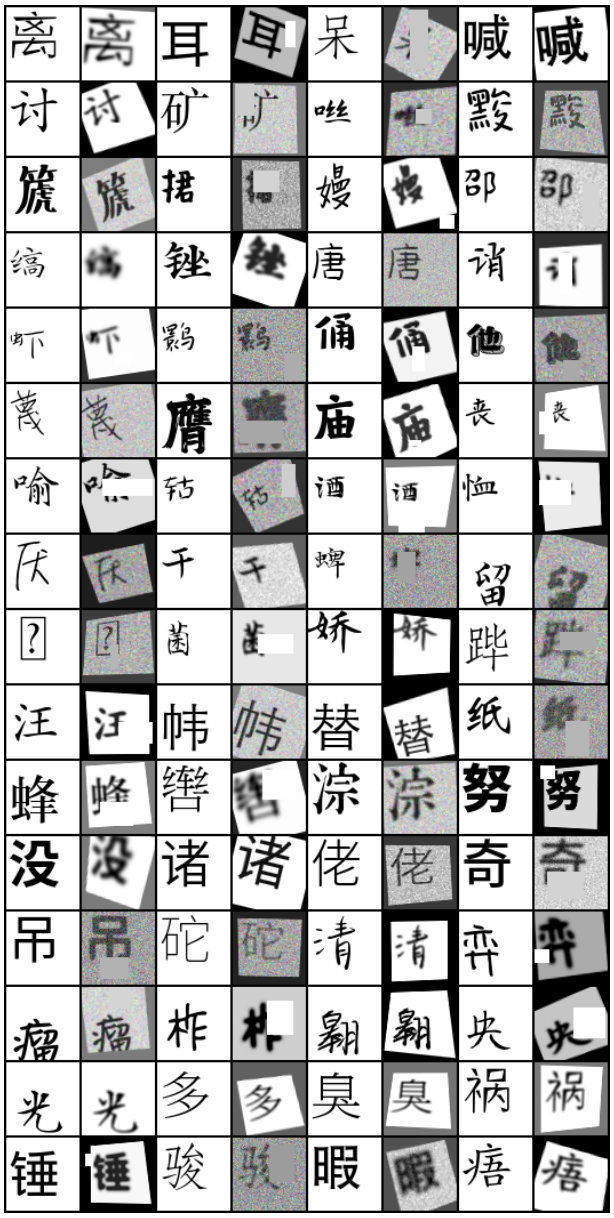


Figure 4.2: Transformations of the images

- Dai Viet Su Ky Toan Thu is another historical records written in classical Chinese. It is written by historian Ngo Si Lien and finished in 1479, in the period of the Le Dynasty. Our data consists images of Dai Viet Su Ky Toan Thu in printed version and hand-written version.

We use Differential Binarization model [10] to get bounding box for the book images. Then, we crop the characters and make them as input to our experiment, the cropped characters for each book can be found in figure 4.3. From figure 4.3, we can see that most characters are correctly cropped, however, in Dai Nam Quoc Su Dien Ca or Truy Mon Canh Huan, due to the nature of the books (characters are written and printed too close to each other), many of the bounding boxes contain 2 or more characters, or the characters do not reside in the center of the boxes. However, since in reality, cases

where human make mistakes labelling are always possible, so we still leave the cropped characters as they are. However, later it still turns out that our approach still produces decent results on this kind of dataset.

As stated, most of the data we have at hand are unlabeled. To obtain labels for any meaningful learning processes, we employed an off-the-shelf OCR tool for Chinese characters recognition. The resulting labels acquired by this way are messy and can be detrimental to the training process if used directly: they include gibberish texts, undetected characters and false predictions. We alleviated this issue by applying some data cleaning techniques to labels. We removed labels that contain non-Chinese characters, such as special symbols and numbers, we also discarded labels whose length is different from 1 since some images contain 2 or more characters. After removal, our dataset has a total of 49k characters as well as their labels. However, to maintain the class balance in the support set and the query set, wehave selected 100 most common character classes from all books only. For each class of characters, we sample out 80 characters, in which 20 is for the support set and 60 is for the query set. The distribution of character in 7 Han Nom books are shown in table 4.1

Table 4.1: Book distribution of the selected characters

Book	Selected characters	Number of Classes
Dai Viet Su Ky Toan Thu Printed	662	81
Dai Viet Su Ky Toan Thu Handwritten	691	83
Truy Mon Canh Huan	814	97
Khoa Hu Luc	1720	99
Dai Nam Thuc Luc Tien Bien	432	75
Dai Nam Quoc Su Dien Ca	130	47
Phap Hoa De Cuong	3551	100

The setting is similar to a setting where users could input a small amount of labeled samples which then can be used to train a classifier to recognize the other images of the label but on a larger scale, this user-provided dataset is usually small and can contain misleading labels.

4.2 Pretrain and evaluate on artificial data

For embedding representation learning, we first train the embedding model on the artificial data generated from 25 fonts and a corpus of Chinese texts consists of 6k characters, making the total of 150k characters in the pre-trained dataset, we augmented the data with extensive transformations. Two different models are trained with Triplet loss and Manifold Mix-up as discussed in the previous section and serves as the initial weights for later fine-tuning on actual data.

After training the embeddings by different losses, 200 novel classes from the artificial

Figure 4.3: Illustration of uncurated characters on 7 books

dataset are sampled for the testing purpose. In this phase, we separates 25 fonts in each class into support and query set, the support set contains 5 fonts and the query set contains 20 fonts of each class.

For the model trained with Triplet loss, werun the query images through the model and compare the distance from them to the support images' embeddings. Top k closest labels are then chosen for prediction and accuracy calculation. L2 distance is employed for this task. The triplet model tested on fonts dataset gives promising results when it achieves 89.4% for top 1 closest label and 97.7% for top 5 closest labels.

For the model trained with Manifold Mixup loss, weuse both K Nearest Neighbors and PT-MAP to select the best character class for prediction. For both techniques, Top 1 and Top 5 closest labels are also selected for accuracy selection. The K Nearest Neighbors technique gives around 95.9% for top 1 and 97.6% for top 5, while PT-MAP techniques give similar result of around 95.5% for top 1 and 96.5% for top 5.

To better understand the insights of the learned feature space in the this step, we conducted a small visualization on the pretrained model (with Triplet loss) to gain some insight of the embedding space. We picked randomly 12 characters on the book dataset and generated these characters with fonts that we used to pretrain the model. The embedding of these characters, both fonts and books, are then processed by projecting them with tSNE on a 2D plane, the result is shown in figure 4.4. We observed that the model learned good representations of the generated characters as they are nicely separated on the projected plane, but it also indicates that this space does not work very well with actual data as all the actual 12 characters from books are grouped into the same cluster. This suggests that we need an additional fine tuning step later to improve the performance on the actual book data.

4.3 Fine-tune and evaluate on Han-Nom books dataset

In the next step, the book dataset is used. In order to get the ground truth labels for learning, we used Easy OCR to extract raw predictions and clean them to make it better suit for training purposes. For label cleaning, we only kept the labels that satisfied some predefined conditions that we thought might help removing irrelevant and misleading labels, such as predicted string length, occurrence of special characters, bounding box ratio and so on. After this step, the total number of characters we acquired from all books is around 49k. Among them, we picked images which belongs to 100 most common characters and splitted them into support and query sets. The support set contains 2000 characters' images, belonging to 100 classes, each class has 20 images in the support set. The query set contains 6000 characters' images, belonging to 100 classes, each class has 60 images in the query set. The distribution of images in the support set and in the query set are visualized in figure 4.5.

After separating the support set and the query set, two models using Triplet loss and Manifold Mix-up Loss are fine-tuned on the support set using the same loss as when

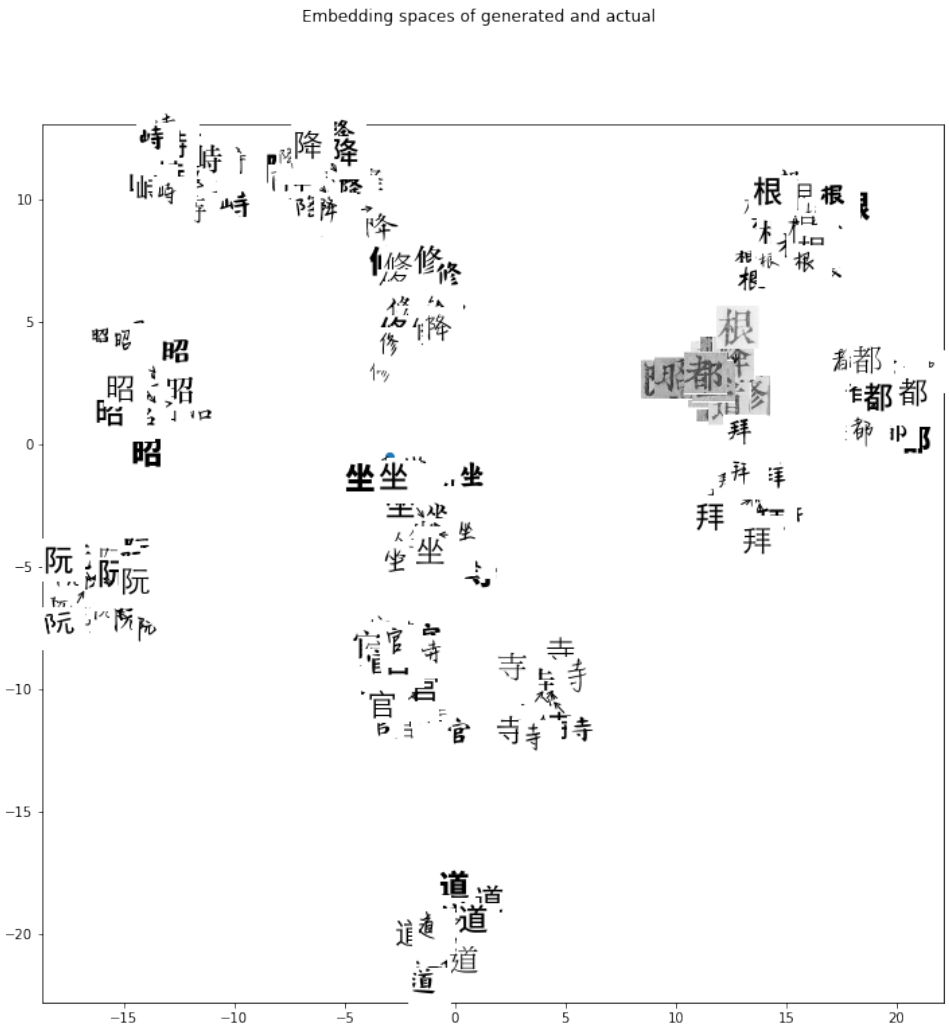


Figure 4.4: Visualization of the embedding space before the training step on mixed data. All 12 generated characters from fonts are seperated on different groups but actual characters all go to the same group. This suggests the need of an additional fine tuing step.

they are trained with fonts. Then, the query images are passed through the fine-tuned models to produce embeddings. The query embeddings are then compared with the support embeddings by different strategies. With triplet model, k Nearest Neighbors are employed to pick embeddings from the support set which are the closest to the target embeddings. With model trained using Manifold Mix-up loss, both k Nearest Neighbors and PT-MAP are used to match the embeddings. The details of K Nearest Neighbors and PT-MAP algorithm are described in Chapter 3.

4.4 Results

As explained in the previous sections of Chapter 4, my experiment includes the following steps:

- Pre-train embeddings model on artificial fonts dataset. The artificial dataset has

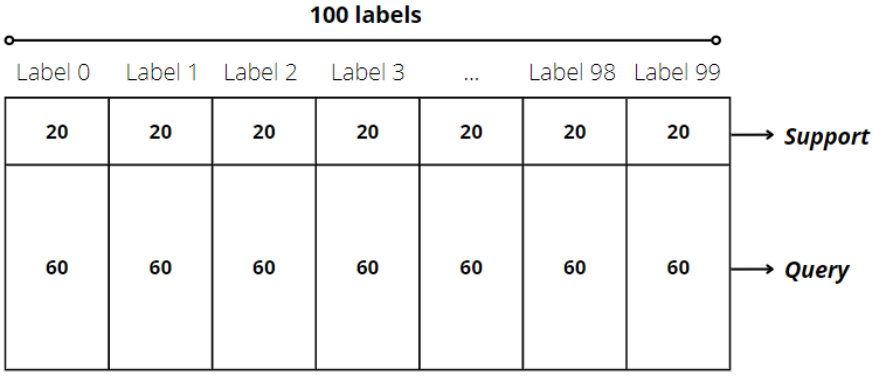


Figure 4.5: Support, query set split from books

6000 Chinese classes with 25 elements for each class.

- Test on 200 novel classes from the artificial fonts dataset (no fine-tune)
- Fine-tune and evaluate on 100 classes of characters from book dataset, each class has 80 elements, 20 for support set and 60 for query set.

The test results on 200 novel characters in artificial dataset are shown in table 4.2. The results show that the models have adapted well and are able to clearly separate the representations for different classes, even for classes that it has never seen before (figure 4.4). However, when directly applied to another dataset with different characteristics like the book dataset, further fine-tuning step are needed as explained in section 4.2.

Table 4.2: Experiment results on font dataset

Type of Loss	Comparison technique	Top 1	Top 5
Triplet	KNN	89.4	97.7
Manifold Mix-up	KNN	95.9	97.6
Manifold Mix-up	PT-MAP	95.5	96.5

After fine-tuning on the book dataset, the models trained with Triplet loss and Manifold Mix-up loss return results with high accuracy. We fine-tune the Triplet model on the book dataset for up to 50 epochs and fine-tune with Manifold Mix-up for 10 epochs only. To prove that my approaches outperform traditional machine learning methods, we conduct another experiment with Resnet50 using only limited support and query sets from the 7 books. The experiment results on the book dataset are shown in table 4.3. We use number to denote different books as followed: 0 is Phap Hoa De Cuong, 1 is Truy Mon Canh Huan, 2 is KhoaHuLuc, 3 is Dai Nam Quoc Su Dien Ca, 4 is Dai Nam Thuc Luc Tien Bien, 5 is Dai Viet Su Ky Toan Thu (printed), 6 is Dai Viet Su Ky Toan Thu (hand-written)

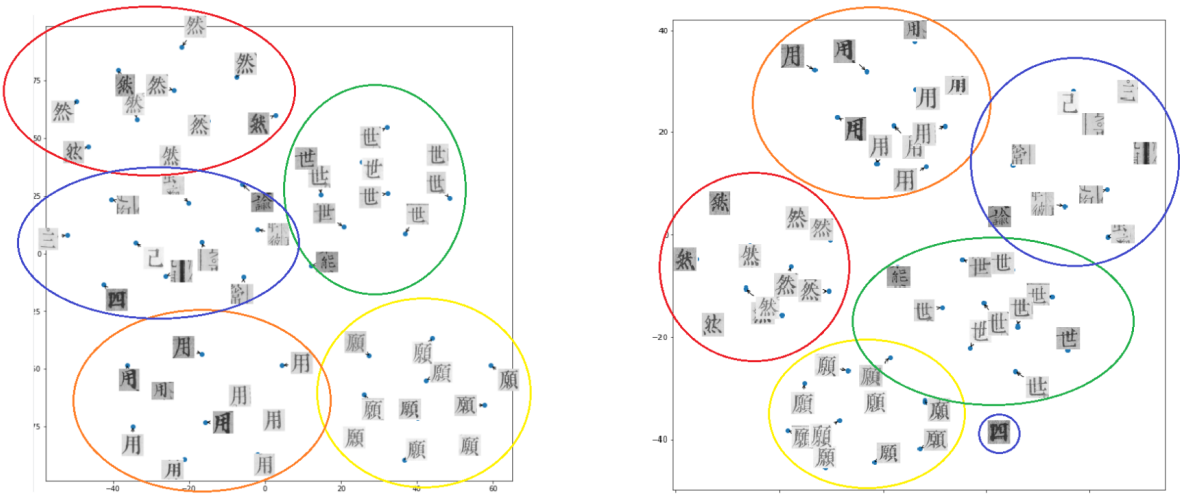
Table 4.3: Top-1 and Top-5 accuracy results in each book with our proposed methods

Book	**Resnet50**		**Triplet + KNN**		**MM + KNN**		**MM + PT-MAP**	
	Top 1	**Top 5**	**Top 1**	**Top 5**	**Top 1**	**Top 5**	**Top 1**	**Top 5**
0	93.6%	98.0%	98.6%	98.8%	99.0%	99.1%	98.9%	99.8%
1	67.4%	83.7%	80.9%	82.4%	83.7%	90.0%	82.7%	94.6%
2	90.0%	97.0%	96.7%	97.0%	97.4%	97.8%	97.4%	99.2%
3	15.8%	35.6%	34.7%	44.6%	38.6%	55.5%	44.6%	73.3%
4	80.2%	88.0%	90.4%	91.3%	91.9%	93.7%	91.3%	96.7%
5	67.8%	82.9%	83.5%	86.1%	86.1%	90.0%	86.5%	95.1%
6	51.9%	71.6%	73.7%	76.9%	78.2%	84.5%	75.9%	92.2%
Overall	82.5%	91.2%	91.4%	92.5%	92.8%	94.8%	92.6%	97.5%

The results table show the performance of different pretrain models as well as embeddings matching techniques on the book dataset and compare them to Resnet 50. All the of fine-tuned models with representation learning techniques achieve accuracy scores greater than 90%. This means that with a few fine-tuning epochs, the representations learning models have efficiently separate the classes by their embeddings. To validate this, we also plot the embeddings of images from 5 random characters (from the query set) after going through models with Triplet loss and Manifold Mix-up Loss (figure 4.6).

3 observations can be made on figure 4.6. First, in both embedding spaces, images from the same class tend to concentrate in one place and separate from the other classes, hence simple matching techniques like K Nearest Neighbor can already produce decent results. Moreover, from the visualization of the embedding space, we can observe that the Manifold Mix-up model's embedding space looks much more concentrated than embedding space of Triplet model, which can also explain the superiority of model trained by Manifold Mix-up loss when compared to model trained by Triplet loss. Second, we can see that the blue ellipses from both embedding space contain mistaken localized characters, which we still allow to be in my dataset to make the experiment closer to real life scenario. The mistaken localized characters to some extent can affect the models' learning procedure. But despite of that, the models still learn good representations, as seen in the visualization in figure 4.6. Third, if we compare between embedding spaces on the book dataset in figure 4.6 and the font dataset in figure 4.4, the fonts' embeddings are clearly more concentrated, thus results in higher accuracy in the font dataset compared to the book dataset.

Another observation is that for top 1 accuracy, models using representation learning techniques surpass traditional resnet50 by 10 to 15% (600 to 900 more correct characters) with less training (10 and 50 epochs compared to Resnet 50's 100 epochs). However, the accuracy of Resnet 50 model on Han Nom book is already acceptable. That raises the question of whether or not we need to gather artificial data and pre-train the representation learning models. In this experiment, my approach beats traditional learning methods by 600 to 900 images. But in reality, there are hundreds of Han Nom documents have not been labelled, and 10% accuracy in my experiment is much larger in reality, 600 to



Embeddings by Triplet Loss Model *Embeddings by Manifold Mix-up Loss Model*

Figure 4.6: Visualization of the embedding spaces after fine-tuning on the support set of 7 Han Nom books. We pick out 5 random classes of character, then sample 10 images in the query set per class (which means the models have never trained on it) and plot out their embeddings on the 2D planes. Ellipses of different colors represent different classes. The 2 figures present 2 embedding spaces produced by 2 models trained by Triplet and Manifold Mix-up loss on the same set of images.

900 images difference can become thousands. As a result, more human labor is needed to re-label the wrong one. Moreover, since Resnet 50 is a traditional machine learning model, it can only classify between a fixed number of class. When images from new class appear, it will have to retrain the model. My representation learning models, on the other hand, learn a vector to represent characteristics of an image, so when a new label appear, we do not have to retrain the model but only need to output embeddings from the existed model, then compare the embeddings with other images.

Next, we will look at some wrong predictions by the models. Figure 4.7 shows typical examples where all models return wrong results. The total number of wrong predictions account for about 2% of all query sets. From the figure, we can observe that most of the wrong predictions here are wrongly labelled, or the image contains multiple character, which confuses EasyOCR in the labelling process. Only a small portion of the wrong predictions from all models are correctly labelled. So if we have a more thorough process of labelling the data, the result can have been even better.

In conclusion, in this chapter, we have presented my experiment pipeline and show the experiment's results. Representation learning methods have shown its promising capability when achieving over 97% over a typical Han Nom documents dataset. The approach also opens up a new way for us to both efficiently label large amount of Han Nom dataset with less human labor and in the mean time receiving recommended labels from users without having to retrain the model like in traditional machine learning.

Figure 4.7: Wrong predictions by all models (split by book)

CHAPTER 6. CONCLUSION AND FUTURE WORK

In this report, we have proposed an approach to deal with the currently large amount of unlabelled Han Nom dataset. We make use of representation learning techniques for characters classification. Both models can easily adapt to new dataset given a few sample.

The carried out experiments show promising results on the proposed the method. The method achieves high results on real world data with limited training data. From the original data of seven Han Nom books, we make use of the open source OCR to localize and pre-label the characters. Then we clean the labelled images, filter and split them in accordance to few-shot setting for the characters classification task. In classification tasks, embedding learning gives good results when the Triplet models give top 5 accuracy up to around 93% on average although being trained on a small amount of data, and the models trained with Manifold Mix-up loss give top 5 accuracy up to around 98% on average.

In addition to the proposed learning method, we also built a web-application that assists users on storing and analyzing Han Nom documents. The initial application has basic use cases that allows users to login, upload books and extract character boxes as well as their suggested labels. In the future, we will focus on extending the application's new features, for example to record the corrected labels from users and add new types of characters to the database.

Given the promising results of our approach, the work could be extended in several directions. Firstly, the dataset could be added with more character and books. Our approach allows the usage of unlabelled data, which is an advantage to the traditional approach. The second one is the direction of using domain adaption methods which consider each book is a specific domain. The embedding learning process are designed to adapt from well-known domain to the new one.

Our work is going to be submitted as a paper with title: *Learning Embeddings for Recognizing Han-Nom Characters in Vietnamese Historical Books*

Bibliography

[1] The essential guide to neural network architectures. https://www.mongodb.com/. Accessed: 2022-01-03.

[2] Under-fitting and over-fittign in machine learning. https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/. Accessed: 2022-01-03.

[3] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

[4] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] Yuqing Hu, Vincent Gripon, and Stéphane Pateux. Leveraging the feature distribution in transfer-based few-shot learning. In *International Conference on Artificial Neural Networks*, pages 487–499. Springer, 2021.

[7] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[8] Fumitaka Kimura, Kenji Takashina, Shinji Tsuruoka, and Yasuji Miyake. Modified quadratic discriminant functions and the application to chinese character recognition. *IEEE transactions on pattern analysis and machine intelligence*, (1):149–153, 1987.

[9] Neil D Lawrence and John C Platt. Learning to learn with the informative vector machine. In *Proceedings of the twenty-first international conference on Machine learning*, page 65, 2004.

[10] Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai. Real-time scene text detection with differentiable binarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11474–11481, 2020.

[11] Moshe Lichtenstein, Prasanna Sattigeri, Rogerio Feris, Raja Giryes, and Leonid Karlinsky. Tafssl: Task-adaptive feature sub-space learning for few-shot classification. In *European Conference on Computer Vision*, pages 522–539. Springer, 2020.

[12] Cheng-Lin Liu. Normalization-cooperated gradient feature extraction for handwritten character recognition. *IEEE transactions on Pattern Analysis and machine intelligence*, 29(8):1465–1469, 2007.

[13] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[14] Puneet Mangla, Nupur Kumari, Abhishek Sinha, Mayank Singh, Balaji Krishnamurthy, and Vineeth N Balasubramanian. Charting the right manifold: Manifold mixup for few-shot learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2218–2227, 2020.

[15] Kha Cong Nguyen, Cuong Tuan Nguyen, and Masaki Nakagawa. Nom document digitalization by deep convolution neural networks. *Pattern Recognition Letters*, 133:8–16, 2020.

[16] Keiron O'Shea and Ryan Nash. An introduction to convolutional neural networks, 2015.

[17] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[18] Leif E Peterson. K-nearest neighbor. *Scholarpedia*, 4(2):1883, 2009.

[19] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2015.

[20] Sagar Sharma, Simone Sharma, and Anidhya Athaiya. Activation functions in neural networks. *towards data science*, 6(12):310–316, 2017.

[21] Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global, 2010.

[22] Truyen Van Phan, Kha Cong Nguyen, and Masaki Nakagawa. A nom historical document recognition system for digital archiving. *International Journal on Document Analysis and Recognition (IJDAR)*, 19(1):49–64, 2016.

[23] Truyen Van Phan and Masaki Nakagawa. Construction of a text digitization system for nom historical documents. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, pages 65–70, 2014.

[24] Truyen Van Phan, Bilan Zhu, and Masaki Nakagawa. Collecting handwritten nom character patterns from historical document pages. In *2012 10th IAPR International Workshop on Document Analysis Systems*, pages 344–348. IEEE, 2012.

[25] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*, pages 6438–6447. PMLR, 2019.

[26] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

[27] Stefan Wager, Sida Wang, and Percy S Liang. Dropout training as adaptive regularization. *Advances in neural information processing systems*, 26, 2013.

[28] Yan Wang, Wei-Lun Chao, Kilian Q Weinberger, and Laurens van der Maaten. Simpleshot: Revisiting nearest-neighbor classification for few-shot learning. *arXiv preprint arXiv:1911.04623*, 2019.

[29] Yaqing Wang, Quanming Yao, James Kwok, and Lionel M. Ni. Generalizing from a few examples: A survey on few-shot learning, 2020.

[30] Yi Yao and Gianfranco Doretto. Boosting for transfer learning with multiple sources. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 1855–1862. IEEE, 2010.

[31] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.