A study of the ratio of labeled and unlabeled data of semi-supervised learning

Xiao Ji, Zining Fan, Zhuoyue Xing

Abstract

In this project, we would like to explore the ratio of labeled and unlabeled data in semi-supervised learning. We will use MNIST dataset and we put aside some data as test data at the beginning. We will use 1000 images as the labeles data, and change the amount the unlabeled data.

1 Introduction

1.1 Related Work

Deep Learning has shown remarkable success in image recognition[1]. We could achieve image classification with deep convolutional neural networks[2]. However, these models need to be trained with supervised learning which requires a large amount of labeled data[3]. By showing the models only labeled images, we limit ourselves from making use of unlabeled images. To avoid this, we could use self-training using unlabeled data in semi-supervised learning [4, 5] to improve our model such as accuracy and robustness.

1.2 Problem Formulation

The goal of our paper is to study two main problems. First, we will change the ratio of the amount of labeled and unlabelled data in the training set and investigate its impact on test accuracy. Second, we would like to find the impact of noise on the test accuracy by adding various combination of noise to the traditional semi-supervised learning model with fixed radio in the first problem and study the difference of test accuracy when we add different combination of noise.
In this project, we tried these models:

1. No labelled data, which is basically random guess;
2. Gaussian noise and no unlabelled data, which is supervised learning;
3. Gaussian noise and change the ratio of labelled data to unlabelled data.

2 Dataset to be used and method proposed

2.1 MNIST Dataset

The study will be conducted using the MNIST Database of handwritten digit, which has a training set of 60,000 examples, and a test set of 10,000 examples.

In our project, we would like to test the ratio of the labelled and unlabelled data. So, we put aside 10000 test samples at first; then we will select some examples from the training and keep their labels, and ignore the labels of the rest of the training set. We will use these labeled images, together with different number of unlabeled images to train our model, and predict the labels of the test images. Then We will add the unlabeled data and gradually increase the ratio of labeled and unlabeled data.

Conference on Neural Information Processing Systems (NeurIPS), .

2.2 Ladder Network

The model we used in our project is ladder network[6]. The basic structure of ladder network is shown as follows:

1. A feedforward model as encoder which serves supervised learning
2. A decoder which could invert the mapping of each layer serving unsupervised learning
3. Train the whole Ladder network by minimizing the sum of all the cost function terms

After building our network, we would like to train our model using labeled and unlabeled data by optimization which is semi-supervised learning.

Algorithm 1

Ladder Network

Require $x(n)$
Corrupted encoder and classifier
$\tilde{\mathbf{h}}^{(0)} \leftarrow \tilde{\mathbf{z}}^{(0)} \leftarrow \mathbf{x}(n) + noise$
for $l = 1$ to L **do**
  $\tilde{\mathbf{z}}^{(l)} \leftarrow \text{batchnorm}\left(\mathbf{W}^{(l)}\tilde{\mathbf{h}}^{(l-1)}\right) + \text{noise}$
  $\tilde{\mathbf{h}}^{(l)} \leftarrow \text{activation}\left(\gamma^{(l)} \odot \left(\tilde{\mathbf{z}}^{(l)} + \beta^{(l)}\right)\right)$
end for
$P(\tilde{\mathbf{y}}|\mathbf{x}) \leftarrow \tilde{\mathbf{h}}^{(L)}$
Clean encoder (for denoising targets)
$\mathbf{h}^{(0)} \leftarrow \mathbf{z}^{(0)} \leftarrow \mathbf{x}(n)$
for $l = 1$ to L **do**
  $\mathbf{z}_{\text{pre}} \leftarrow \mathbf{W}^{(l)}\mathbf{h}^{(l-1)}$
  $\boldsymbol{\mu}^{(l)} \leftarrow \text{batchmean}\left(\mathbf{z}_{\text{pre}}^{(l)}\right)$
  $\boldsymbol{\sigma}^{(l)} \leftarrow \text{bat chstd}\left(\mathbf{z}_{\text{pre}}^{(l)}\right)$
  $\mathbf{z}^{(l)} \leftarrow \text{batchnorm}\left(\mathbf{z}_{\text{pre}}^{(l)}\right.$
  $\mathbf{h}^{(l)} \leftarrow \text{activation}\left(\boldsymbol{\gamma}^{(l)} \odot \left(\mathbf{z}^{(l)} + \boldsymbol{\beta}^{(l)}\right)\right)$
end for
Decoder and denoising
for $1 = \text{L}$ to 0 **do**
  if $1 = \text{L}$ **then**
    $\mathbf{u}^{(L)} \leftarrow \text{batchnorm}\left(\tilde{\mathbf{h}}^{(L)}\right)$
  else
    $\mathbf{u}^{(l)} \leftarrow \text{batchnorm}\left(\mathbf{V}^{(l+1)}\hat{\mathbf{z}}^{(l+1)}\right)$
  end if
  $\forall i : \hat{z}_i^{(l)} \leftarrow g\left(\tilde{z}_i^{(l)}, u_i^{(l)}\right)$
  $\forall i : \hat{z}_{i,\text{BN}}^{(l)} \leftarrow \frac{\hat{z}_i^{(l)} - \mu_i^{(l)}}{\sigma_i^{(l)}}$
end for
Cost function C for training
$\text{C} \leftarrow 0$
if $t(n)$ **then**
  $\text{C} \leftarrow -\log P(\tilde{\mathbf{y}} = t(n)|\mathbf{x}(n))$
end if
$\text{C} \leftarrow \text{C} + \sum_{l=0}^{L} \lambda_l \left\|\mathbf{z}^{(l)} - \hat{\mathbf{z}}_{\text{BN}}^{(l)}\right\|^2$ =0

2.3 Change the Ratio

After building the ladder network model, we will use 1000 labeled data and train our model repeatedly by changing the ratio of labeled data to unlabeled data from 1:1 to 1:500.

3 Results

We used the mnist dataset from keras.datasets package to train and test our model. We have tried the original code of efficient-net but it is hard for us to fully understand it in a short time. Instead, we built a ladder net model to study the ratio of labeled and unlabeled data first.

3.1 No Noise and No Labelled Data

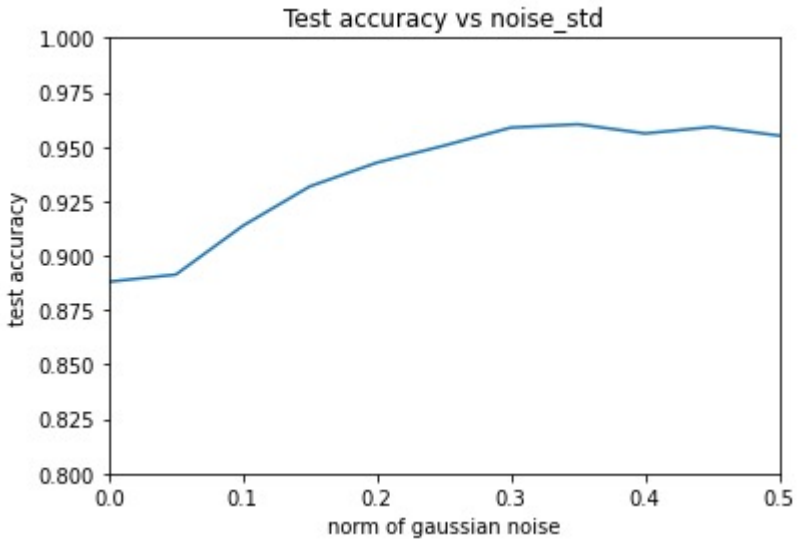

Figure 1:

3.2 Change the ratio

To optimize the radio of unlabeled data to all training data in our model, we randomly selected 1000 images as labeled data and changed the size of unlabeled data. The sizes of unlabeled data we used in our model were 1000, 2000, 5000, 10000, 20000 and 50000. We trained all models for 3 epochs. Figure 2 shows the test accuracy of different models.

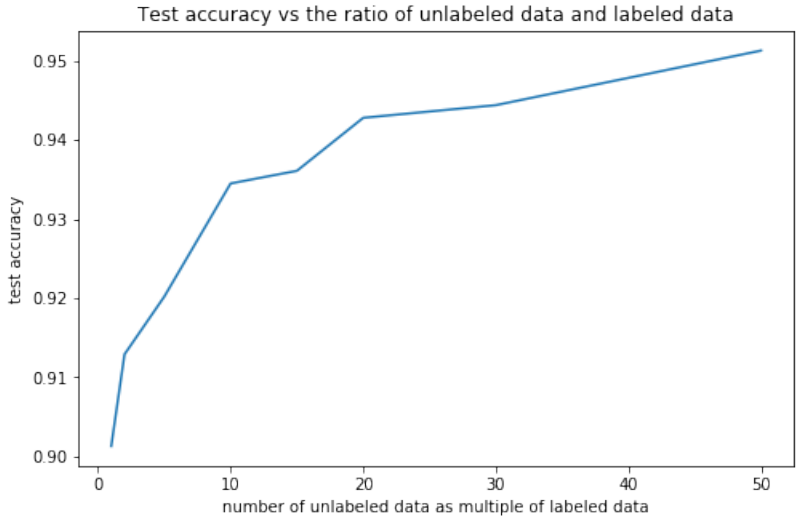

Figure 2:

In Figure 2, we can see that the test accuracy reaches 90% even when we used only 1000 labeled images and 1000 unlabeled images. Also, the test accuracy keeps increasing as the size of unlabeled dataset goes up. Thus, we decided to use only 500 labeled data in order to explore a larger range of the labeled and unlabeled data ratio.

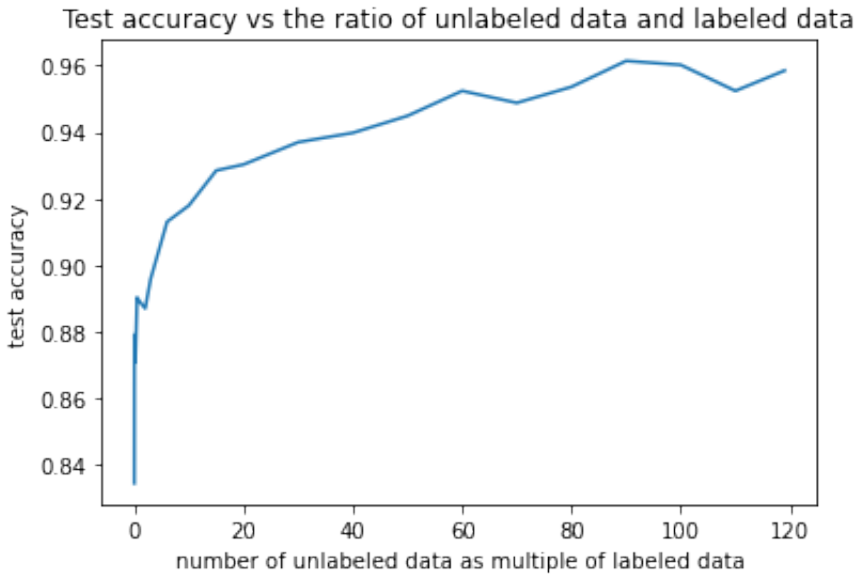

Figure 3:

Figure 3 shows the test accuracy of using the combination of 500 labeled and 1 to 59500 unlabeled data, which provides the range of labeled/unlabeled ratio from 500:1 to 1:119. We could not set the number of unlabeled data to 0 due to the setting of LadderNet model. We ran 10 epoches for each model.

The starting test accuracy trained by 500 labeled data and 1 unlabeled data is 83%. The test accuracy increased rapidly to 93% when labeled/unlabeled ratio increases to 1:20. This improvement makes sense because we added an unsupervised learning targets on every layer which captures more relevant details than a supervised learning model on the top layer. As we included more unlabeled data, the test accuracy keeps increasing but at a lower speed. The reason may be that the 20 x 500 additional unlabeled images have already given us many important details, so adding more unlabeled images won't provide much valuable information. The test accuracy hits the highest point when the labeled/unlabeled ratio is 1:90, with a value of 96%.

The Github repo link of our project is github.com/ZiningFan00/semi-supervised-learning and our code is there.

4 Future goals

1. We will try to find a better number of epochs and the size of labeled data set.
2. We will investigate the impact of noise to the test accuracy by adding input noise only, model noise only and the combination of the two.
3. The net model we use now is relatively simple. We will improve our net model to improve the test accuracy of our model in the future.

References

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems*. 2012, pp. 1097–1105.

[2] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[3] Mingxing Tan and Quoc V Le. "Efficientnet: Rethinking model scaling for convolutional neural networks". In: *arXiv preprint arXiv:1905.11946* (2019).

[4] Qizhe Xie et al. "Self-training with Noisy Student improves ImageNet classification". In: *arXiv preprint arXiv:1911.04252* (2019).

[5] Bruno Lecouat et al. *Semi-Supervised Deep Learning for Abnormality Classification in Retinal Images*. 2018. arXiv: `1812.07832 [cs.CV]`.

[6] Antti Rasmus et al. "Semi-supervised learning with ladder networks". In: *Advances in neural information processing systems*. 2015, pp. 3546–3554.