PROCEEDINGS OF SCIENCE

The Application of Autoencoder in Classification of the Eye Movement Data

Mengjie Zhang [1a],Mi Li[2,3b], Jiankang Sun[c], Shengfu Lu[d]
International WIC Institute, Beijing University of Technology, Beijing 100124, China
Beijing International Collaboration Base on Brain Informatics and Wisdom Services, Beijing 100124, China
Beijing Key Laboratory of MRI and Brain Informatics, Beijing 100053, China
E-mail: [a]S201202126@emails.bjut.edu.cn;[b]limi@bjut.edu.cn
[c]ytusunjiankang@emails.bjut.edu.cn;[d]lusf@bjut.edu.cn

Ning Zhong[1]
International WIC Institute, Beijing University of Technology, Beijing 100124, China
Beijing International Collaboration Base on Brain Informatics and Wisdom Services, Beijing 100124, China
Beijing Key Laboratory of MRI and Brain Informatics, Beijing 100053, China
Maebashi Institute of Technology Maebashi-City 371-0816, Japan
E-mail: zhong@maebashi-it.ac.jp

Deep learning well demonstrates its potential in learning latent feature representations, and has been applied in the fields of speech recognition, image identification and information retrieval. Deep learning architecture is composed of multilayer non-linear units, each low layer's output as a input of higher layer, and can learn high-order feature representations which contain many structural information from a large number of data. Deep learning is a good way to extract features from original data. Web page is an important human-machine interface. Identify users' visual behavior on Web pages will promote human-machine interaction. This paper apply stacked auto-encoders (SAE) with logistic regression to build classification model. This model effectively solves the problem of identifying users' working state of visual search and visual browse on Web pages. The experiment shows that this model outperforms other Single model such as SVM and logistic regression and achieves the accuracy of 90.32%. Further, we embed adaboost algorithm to improve recognition accuracy and precision.

ISCC 2015
18-19, December, 2015
Guangzhou, China

[1]Speaker

[2]This study is supported by the International Science & Technology Cooperation Program of China (No. 2013DFA32180), the 973 Program (No. 2014CB744600), National Natural Science Foundation of China (No. 61420106005), the Beijing Natural Science Foundation (No. 4132023), the Beijing Outstanding Talent Training Foundation (No. 2014000020124G039).
[3]Corresponding Author

© Copyright owned by the author(s) under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0).

PoS(ISCC2015)001

1.Introduction

Leaning feature representations from deep neural networks has been a popular method recent years. It has gained numerous success in the fields such as speech recognition [1], image identification[2]and information retrieval[3]. Qirong Mao propose to learn affect-salient features for speech emotion recognition using convolutional neural networks. It is confirmed that the features extracted by this model for classification are better than several well-established speech emotion features[4]. Eswaran found that reconstruction error of stacked auto-encoder pre-trained by RBM is better than the traditional autoencoder with RBM and the stacked autoencoder without RBM[5]. Salakhutdinov et al. put forward a semantic hash method which utilizes features learned by deep neural network for information retrieval[6]. So, deep learning can find appropriate features.

Eye tracking technique has been applied in human-machine interaction, which can record users' eye movement data in real-time[7]. Andreas et al. use SVM classfiy users' visual behavior, whose model reach up to 76.1% on average[8]. Jella et al. used mobile eye-tracking technology to explore attentional differences between goal-directed search and exploratory search, and built a model to learn attentional discriminative features, the classification accuracy is 77% [9]. However, the existing methods are mainly based on hand-crafted features. So we are not sure what visual behavior features can result in satisfying classification results.

Web page is an important human-machine interface. People's visual behavior on Web pages is complex and difficult to select features for pattern recognition. Inspired by feature representation of deep learning, this paper employs stacked auto-encoder to extract feature representations from origin data when people participate in visual search and visual browse tasks on Web pages, followed by supervised logistic regression to classify the two cognitive state. The model achieved 90.32% accuracy. Comparing classification results of logistic regression with stacked auto-encoders (AE_LR), logistic regression, and SVM with RBF kernel (RBF_SVM), pre-trained features data by stacked autoencoders can lead to better classification result.

2. Approaches

2.1 Stacked Autoencoder and Logistic Regression

Stacked auto-encoder is a deep learning neural network built with multiple layers of auto-encoders, in which the output of each layer is connected to the input layer of the next layer. SAE learning is based on a greedy layer-wise unsupervised training, which was put forward by Hinton in 2006[10]. It solves the problem that it is hard to train auto-encoders with many hidden layers due to large or small initial weights. Auto-encoder consists of input layer, hidden layer and output layer. The hidden layer can learn data representations from input layer, we can apply these representations for classification tasks. Autoencoder consists of two parts: a coder and a decoder. The encoder is a function F that maps an input $x \in \mathbb{R}^D$ to a hidden representation $z \in \mathbb{R}^K$. It has the form as

$$z = F(x) = S_f(Wx + b) \tag{2.1}$$

PoS(ISCC2015)001

where $W \in \mathbb{R}^{K \times D}$ is a weight matrix, $b \in \mathbb{R}^{K}$ is a hidden bias vector, and S_f is an activation function, which is shown as follows:

$$sigmoid(t) = 1/(e^{-t}+1) \tag{2.2}$$

The decoder function G maps the hidden representation z back to a reconstruction x':

$$x' = G(z) = S_g(W' z + b') \tag{2.3}$$

Where $W' \in \mathbb{R}^{D \times K}$ is a weight matrix, $b' \in \mathbb{R}^{D}$ is a bias, and S_g is a decoder's activation function, sigmoid.

The objective of a classic autoencoder is to minimize the reconstruction error on a training set. The cost function adopted in this paper is Eq.(2.4)

$$min(J) = min(J_1 + J_2) \tag{2.4}$$

Where J is the cost function of AE model. The squared reconstruction error, J_1 ,can be denoted as Eq.(2.5)

$$J_1 = (1/2)\sum_{i=1}^{m} \| x_i - x_i' \|^2 \tag{2.5}$$

J_2 is a weight decay term. It can be represented as Eq.(2.6)

$$J_2 = (1/2)\lambda \left(\|W\|_F^2 + \|W'\|_F^2 \right) \tag{2.6}$$

Where $\|\bullet\|_F$ is the F-norm of matrix and λ is the weights decay parameter.

The gradient values of parameters are calculated with back-propagation algorithm. The limited-memory Broyden-Fletcher-Goldfarb-Shanno(L-BFGS) algorithm which can often be much faster than the gradient descent algorithm can be applied to constantly update the learning rate and gradient values of the parameters so as to find the optimal parameter $\phi \in \{W, b, W', b'\}$. Once a stack of auto-encoders are built according to greedy layer-wise training, the top level data representation can be used as input to a supervised learning algorithm. These algorithms include logistic regression or SVM, etc. The supervised algorithm adopted in this paper is logistic regression which has widely applied to two class classification tasks. The objective of logistic regression is to learn the mapping function from data to label, and then finds the optimal parameter θ through L-BFGS algorithm to minimize the cost function on a training set. The cost function of logistic regression adopted in this paper is Eq.(2.7)

$$J_{lr}(\theta) = -\left(\sum_{i=1}^{m} y_i \log \overline{y}_i + (1 - y_i) \log (1 - \overline{y}_i) \right) \tag{2.7}$$

Where yi is the true label vector of every eye movement data sample, and $\bar{y}i$ is the prediction label vector of that data. In our model shown in Figure1, We add a logistic regression layer on top of the autoencoders to yield a deep neural network to supervised learning.Then, based on the

PoS(ISCC2015)001

labelled data, all parameters of the whole neural network are fine-tuned using a backpropagation technique. This method is called as fine-tuning which has been proved to further improve the recognition accuracy[11].

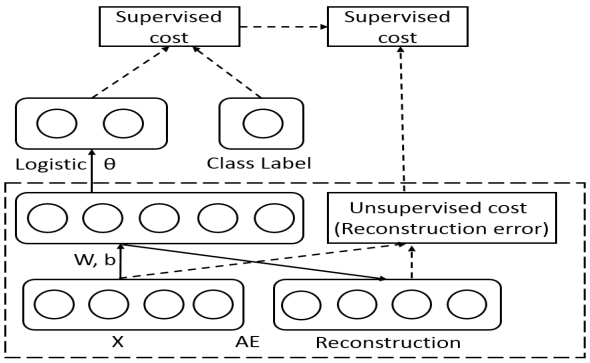

Figure 1: Structure of AE_LR model. The paper apply two layers of auto-encoder to extract features as input of logistic regression layer and output the class label ultimately

2.2 Adaboost Algorithm

Adaboost algorithm is a ensemble learning method. Each time when training a weak predictor iteratively, the sample weight will be adjusted according to classification error rate. The correct weight increase and the error weight decrease. In this way, the misclassification samples will attract more attention in the next iteration. The advantage of adaboost algorithm is that it uses the selected training data after weighting, instead of the randomly selected training samples. The best classifier is selected via weight voting mechanism. We apply adaboost algorithm to combine AE_LR weak predictors to improve the classification accuracy and precision, as in Figure 2.

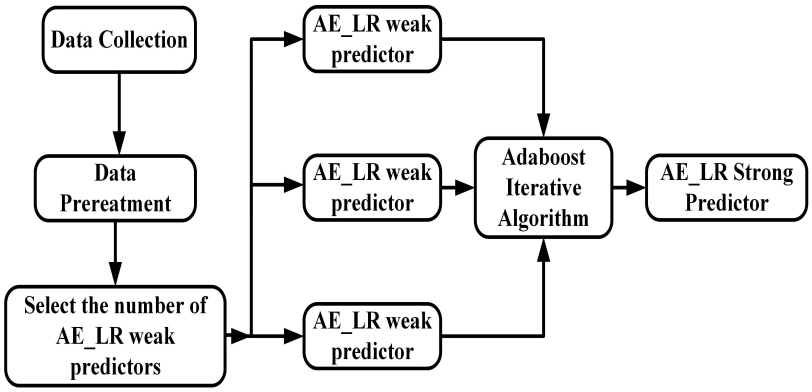

Figure 2: Procedure of AE_LR model based on adaboost algorithm.First, the original data are pretreated. Second, We select five AE_LR models to form a better model through adaboost algorithm

3. Experimental Verification and Interpretation of Result

3.1 Database

This study apply Tobii T120 eye tracker produced in Swedish to record eye movement data of 35 people when they participate in visual search and visual browse tasks on Web pages.

PoS(ISCC2015)001

The paper launch an experiment on ten Web pages about mobile phone, computer and food, etc. The search task is to mine text or image information on Web pages. The browse task is to browse Web pages based on users' preferences and interests. The sampling frequency is 120Hz. Five well-established data of eye movement are selected, including pupil diameter, center distance of gaze point, saccade distance, fixation time and fixation count. Each class has 637 samples. 600 training samples and 674 test samples are adopted in this paper.

3.2 Analysis of Result

The all experiments test on MATLAB R2012b, 4GB, intel i5 PC. Different neural network structure, different number of hidden layers and hidden nodes for each layer are studied. The classification accuracy and test time as the evaluation methods. The classification accuracy is right proportion of total samples. The result, as shown in Table 1, indicates that a model with two hidden layers yield the best performance, compared with 'deeper' models of the same 'width'. Therefore, model structure with two hidden layer is adopted.

Number of layer	Classification accuracy
1	90.72%
2	92.20%
3	90.37%
4	89.28%

Table 1: Recognition accuracy at different layers

Origin data are mapped to a higher dimensional space just like SVM, so as to increase the separability of data. The first layer maps the original data to 3-fold dimensions space. Thus, there are 15 nodes in the first hidden layer. Data compressed and extracted through the second hidden layer. Finally, The logistic regression output the class of users' visual behavior on Web pages. The maximum number of iteration during fine-tuning is 400. At the same time, the weight restraint coefficient is 0.01 during back-propagation. For comparing our algorithms, state-of-the-art classification method—RBF_SVM and Logistic are tested on the same database for comparison. Table 2 and Figure 3 respectively show the classification result and the efficiency of time.

NO	AE_LR	RBF_SVM	Logistic
1	91.17%	87.58%	78.34%
2	89.84%	89.23%	77.27%
3	92.78%	86.42%	78.07%
4	88.77%	85.59%	79.41%
5	89.03%	84.82%	79.42%
x	90.32%	86.72%	79.50%

Table 2: The classification accuracy and average accuracy of three model in five trials

PoS(ISCC2015)001

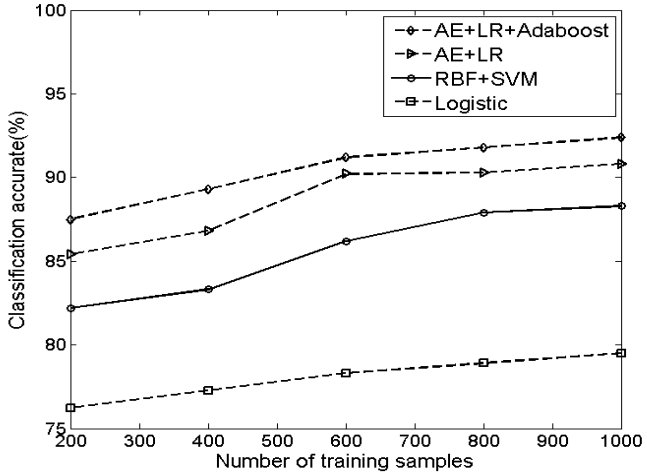

Figure 3: Classification accuracy rate of different methods

Figure 3 shows that AE_LR model outperforms logistic regression and RBF_SVM. Compared with logistic regression, pre-trained feature using stacked autoencoders outperforms original feature data.

Method	Cost time
AE_LR	25ms
RBF_SVM	44ms
Logistic	19ms

Table 3: The efficiency of time of different methods

Conclusions can be drawn from Table 3: although logistic can achieve a good performance, its classification accuracy weaker than AE_LR. As a whole, the efficiency and accuracy of AE_LR can achieve ideal result.

4. Conclusion

The paper proposes a semi-supervised learning method which combines stacked auto-encoders and logistic regression and solves the classification problem of online users' visual behavior, including visual search and visual browse on Web pages. By comparing our model with other models, we find that original data pre-trained by autoencoders proform well. Furthermore, We apply adaboost algorithm to improve AE_LR model's classification accuracy, which suggests adaboost conbined with AE_LR can improve the classification proformance.

References

[1] O.Margarita,L.C.Yann,M.Matthew L.*Synergistic Face Detection And Pose Estimation With Energy-based Models*[J].Journal of Machine Learning Research.8(May 2007),8(1):1197-1215 (2007)

[2] H.Geoffrey,D.Li,Y.Dong,D.George,J.Navdeep,S.Andrew.*Deep Neural Networks For Acoustic Modeling In Speech Recognition:The Shared Views Of Four Research Groups*[J].IEEE Signal Processing Magazine.29(6),82-97(2012)

[3] P.S.Huang,X.D.He,J.F.Gao,L.Deng,A.Alex,H.Larry.*Learning Deep Structured Semantic Models for Web Search using Clickthrough Data*[C].International Conference on Information and Knowledge

PoS(ISCC2015)001

Management, Proceedings.Association for Computing Machinery, United States.pp,2333-2338(2013)

[4] Q.R.Mao,M.Dong,Z.W.Huang,Y.Z.*Zhan.Learning Salient Features for Speech Emotion Recognition Using Convolutional Neural Networks*[J].IEEE Transactions on Multimedia.16(8),2203-2213(2014)

[5] C.C.Tan,C.Eswaran.*Reconstruction Of Handwriten Digit Images Using Autoencoder Neural Networks*[C]. Canadian Conference on Electrical and Computer Engineering,2008.Institute of Electrical and Electronics Engineers Inc,United States.pp,465-469(2008)

[6] S.Ruslan,H.Geoffrey.*Semantic Hashing*[J].International Journal of Approximate Reasoning.50(7),969-978(2009)

[7] Z.C. Feng, M.W.Shen. *Application of Gaze Tracking In Human Computer Interaction*[J].Journal of ZheJiang University (Science Edition) .29(2),226-232(2002)(In Chinese)

[8] B.Andreas,W.JamieA,G.Hans,T.Gerhard.*Eye movement analysis for activity recognition using electrooculography*[J].IEEE Transactions on Pattern Analysis and Machine Intelligence.33(4),741-753(2011)

[9] P.Jella,M.Martin,P.Jascha,P.Thies.*Classification of goal-directed search and exploratory search using mobile eye-tracking*[C].35th International Conference on Information Systems "Building a Better World Through Information Systems", ICIS 2014.Association for Information Systems,New zealand

[10] H.Geoffrey,S.Ruslan.*Reducing the dimensionality of data with neural networks*[J]. Science.313(5786),504-507(2006)

[11] G.Hanlin,T.Nicolas,C.Matthieu,L.Joo-Hwee.*Learning Deep Hierarchical Visual Feature Coding*[J].IEEE Transactions On Neural Networks And Learning Systems.25(12),2212-2225(2014)

PoS(ISCC2015)001