Hindawi
Mobile Information Systems
Volume 2022, Article ID 8230154, 6 pages
https://doi.org/10.1155/2022/8230154

Research Article

A Loss Function Base on Softmax for Expression Recognition

Jin Lu [1] **and Bo Wu** [2]

[1] *Shenzhen Polytechnic, Guangdong Key Laboratory of Big Data Intelligence for Vocational Education, Shenzhen 518055, Guangdong, China*
[2] *Shenzhen Pengcheng Technician College, Guangdong Key Laboratory of Big Data Intelligence for Vocational Education, Shenzhen 518038, Guangdong, China*

Correspondence should be addressed to Bo Wu; wubo@szpt.edu.cn

Received 26 April 2022; Accepted 19 May 2022; Published 6 June 2022

Academic Editor: M. Praveen Kumar Reddy

Benefiting from deep learning, the accuracy of face expression recognition tasks based on convolutional neural networks has been greatly improved. However, the traditional SoftMax activation function lacks the ability to discriminate between classes. To solve this problem, the industry has proposed several activation functions based on softmax, such as A-softmax, LMCL, etc. We investigate the geometric significance of the weights from a fully connected layer and consider the weights as the class centers. By extracting the feature vector of several samples and extending the corresponding means to the weights, the model can develop the ability to recognize custom classes without training, while maintaining the accuracy of the original classification. On the expression task, the original seven-category classification is validated to obtain 97.10% accuracy on the CK+ dataset and 88% accuracy on the custom dataset.

1. Introduction

With the rapid development of deep learning in the past decade, many traditional machine learning tasks have made great progress, and facial expression recognition based on the convolutional neural network has achieved more than ten points on the general validation set. Since AlexNet [1], convolutional neural networks (CNNs) have been widely used in computer vision tasks and CNN can accurately model. A high-dimensional embedding representation is extracted from the input data, and then, a fully connected layer and a softmax activation function are applied to the high-dimensional embedding to minimize the cross entropy between the output of softmax and the real label, and the classification accuracy of the model is improved by iteratively optimizing the objective function.

However, the original softmax activation function has several problems; first, it is sensitive to the modular length of the weight of the fully connected layer and the modular length of the input of the fully connected layer. The longer the module length is, the larger the value of softmax corresponding to the classification output is and the smaller the value of the loss function is, which is easy to cause the model to stop early during training [2]. Second, softmax does not encourage classification between classes (As shown in Figure 1, we can see that softmax only separates different classes, while LMCL and other methods also form interclass gaps). In response to this situation, Wen et al. proposed Center Loss [3]. In the face recognition task, in order to increase the interclass distance of different face IDs and reduce the intraclass distance of the same face ID, based on the original softmax classification, a set of weights are added, which represent the center of each face ID in the space. In each training iteration, not only the original model is optimized by the cross entropy but also the center of the face ID is updated by iterating the set of weights with the Euclidean distance between the corresponding classification data and the classification center as the objective function.

In order to achieve the purpose of reducing the intraclass distance and increasing the interclass distance, A-softmax [4] was proposed by Liu et al., which also normalizes the input of the last fully connected layer on the basis of softmax, normalizes the weight of the fully connected layer, and introduces the angle boundary coefficient m in order to

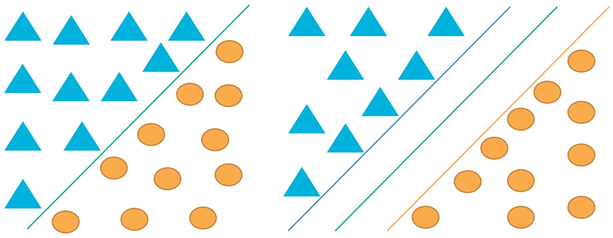

Figure 1: Comparing the boundaries formed by softmax and LMCL.

achieve the goal of increasing the interclass distance and reducing the intra-class distance in the face recognition task. Wang et al. proposed the Large Margin Softmax Loss (LMCL) [5], which is the method proposed by Tencent Artificial Intelligence Laboratory to solve the face recognition task. Similar to A-softmax, LMCL normalizes the input of the last fully connected layer. These methods have effectively improved the original softmax's poor cohesive ability and small interclass spacing.

At the beginning of the 21th century, Ekman and Friesen defined six basic expressions [6]. Since then, the research on expression recognition has been endless. Especially in the last decade, the task of expression recognition has benefited from the rapid development of deep learning and has made great progress. Unlike the traditional expression classification algorithm, it is divided into two parts: feature extraction and feature classification. The expression classification algorithm based on deep learning uses an end-to-end mode to directly obtain expression classification results. With CK+ dataset as the standard, some algorithms based on neural network models [7–9] have achieved more than 95% accuracy. Ding et al. [10] use face recognition datasets to train convolutional networks and the trained convolutional networks to learn facial expression recognition tasks. Ranjan et al. [11] use the basic network of face recognition to realize face smiling face detection on the low-level network and realize different tasks on different network layers.

In the specific task of facial expression recognition, we notice that some specific requirements cannot be achieved based on softmax: first, the facial expression recognition algorithm based on deep learning uses softmax as the activation function of the last connection layer and cross entropy as the target function for training. This method defaults that the classification of all facial expressions has been fixed and will not change. In fact, there may be a lack of expression definition and there may be undefined expression classification. If the inference data contain undefined classification data, using softmax activation, the inference result will be wrong. Second, based on Softmax activation function training, a boundary is formed between classes, but the degree of separation between classes is not emphasized (Figure 1). A small amount of change or noise in the data near the boundary may lead to the change of the classification results, which shows that the recognition results have been shaking back and forth in the continuous sequence recognition task.

Thirdly, the facial expression recognition task has the scene of temporarily recognizing some custom expressions, and the softmax activation method not only needs to retrain the model, At the same time, it is necessary to collect a large number of new custom classification data. Sometimes the cost of collecting and labeling data is very high, and it is not suitable to use the method of collecting data, labeling data, and then retraining.

Lastly, sometimes we need to analyze the relationship between expressions, visualize the distribution of facial features in space, and the boundary between classes formed by softmax activation function is complex. It is not convenient to visualize the distribution of data in the output space of the model. To meet these requirements, we reanalyze the geometric meaning of weights W such as softmax, A-softmax, and LMCL and propose the method of maintaining and updating weights W to realize self-defined classification and recognition.

2. Related Work

The CK+[12] dataset is commonly used in the industry to evaluate the effect of facial expression recognition tasks. CK+ contains 593 sequences of face images, and each sequence contains 10 to 60 frames of expression changes from neutral to fully developed. Among them, 327 sequences are included in 7 categories (anger, contempt, disgust, fear, happiness, sadness, and surprise) and we generally add the last 1 to 3 frames of each sequence to the verification set.

ExpW [13] contains 91793 facial images downloaded from the Internet using a search engine. Each image is manually annotated and classified into seven basic expression categories. We use ExpW as the training set.

RAF-DB [14] contains 29672 real-world face images downloaded from the Internet. The dataset is labeled with crowdsourced patterns. The dataset is divided into two subsets. One subset contains 15339 face images with 7 types of basic expressions, and the other subset contains 11 types of composite face images.

3. Methodology

By analyzing A-softmax, LMCL, and other methods in face recognition, it is found that the weight of the last fully connected layer is abandoned, and the face data only use the previous model for forward reasoning to extract feature vectors. The normalized feature vectors all fall on the hypersphere, and only need to calculate the cosine value with the feature vectors stored in the database. The degree of difference between the two data can be compared.

Inspired by Center Loss's proposal to use weight memory classification centers, we rethink the geometric meaning of the weights of fully connected layers based on A-softmax and LMCL training. We propose to use A-softmax and LMCL loss function training in the task of facial expression recognition, save the weight of the fully

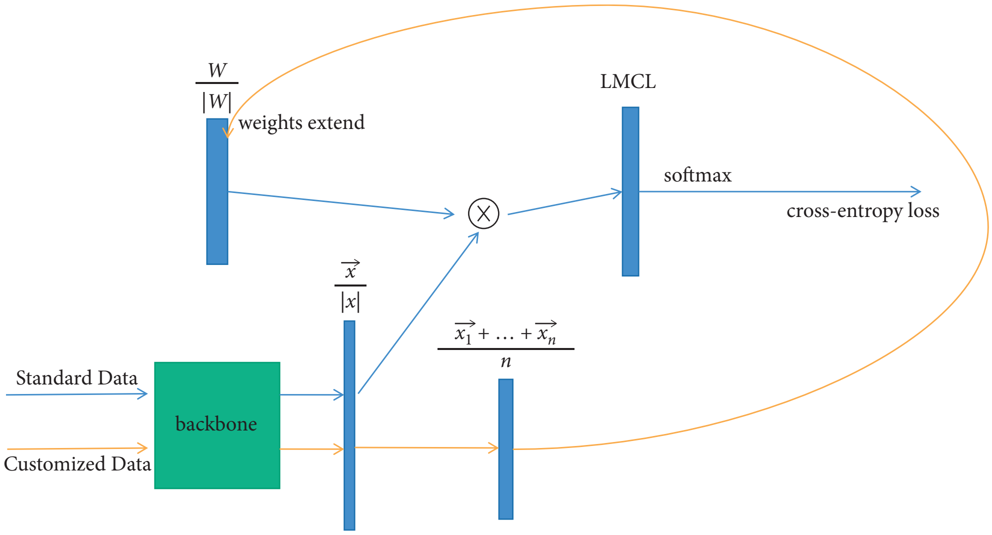

Figure 2: Update taxonomy center flowchart.

connected layer as the classification center in the inference stage, and compare the cosine distance between the extracted feature variables and the classification center O to obtain a classification result.

By retaining the weight of A-softmax and LMCL, the classification results can be obtained only by forward reasoning and softmax activation in the scene where the original seven basic expressions need to be recognized. In the scene where more self-defined expressions need to be recognized, only a plurality of sample data of the self-defined class are used for forward reasoning, and the extracted feature vectors are normalized. The mean value of the normalized feature vector is the class center of the custom classification, the class center is added to the class center weight, and the custom expression can be recognized by using softmax activation.

The softmax loss function is first analyzed and softmax separates the between-class features by maximizing the posterior probability corresponding to the correct label.

The formula is as follows:

$$L = \frac{1}{N}\sum_{i=1}^{N} -\log, p_i = \frac{1}{N}\sum_{i=1}^{1} -\log \frac{e^{f_{yi}}}{\sum_{j=i}^{C} e^{f_i}}. \tag{1}$$

where p_i represents the corresponding posterior probability, N is the total number of training samples, C is the total number of classifications, and f_i represents the output of a fully connected layer:

$$f_i = W_i^T x + B_i. \tag{2}$$

LMCL filed an order $B_i = 0$; W_i and x are normalized, the modified fully-connected layer operator becomes as follows:

$$f_i = \left\|W_i\right\| \|x\| \cos\, \theta_i. \tag{3}$$

That is, the output is the cosine value between the weight and the input sample. The smaller value represents weight x and W_i. The more similar it is, it can be considered that W_i is the center of the classification. We keep the training phase of the model W_i. For regular classification, in the scenario where additional classification is needed, we calculate the class center of the newly added class and update the new class center to W_i :

$$W_j = \sum_{n}^{1} |\vec{x}|, \tag{4}$$

where n is the number of samples used to update the class center; the larger the number, the more accurate the class center, we use $n = 10$; and $\vec{x}$ is the sample feature vector to be calculated. In the inference stage, the maximum value of the classification output is greater than the set threshold of 0.6, which is the corresponding classification, as shown in Figure 2.

Our proposed method not only increases the interclass distance and decreases the intraclass distance but also avoids the frittering of recognition results near the classification border. When for some scenario that requires rapid recognition of custom expressions, our method only needs a few samples of new classes to calculate the class center to obtain the ability to recognize new classes, avoiding the expensive cost of collecting data, labeling data, and also avoiding the cost of a large amount of computer resources and time. Our method allows the inferred features to fall on the hydrosphere, which can be used for expression analysis.

Also, our proposed method can update the center of categories using the new instance of the old category; the new instance data forward inference to obtain the feature vector and update the center of categories using the smooth average.

As can be seen from Table 1, compared to other methods, our proposed method can not only get new knowledge from new instances of original category but can also gain the knowledge of new category and develop the ability to recognize new category, without original training data.

Table 1: Comparative analysis of relative works.

Method	New category	New instance of past category	Do not need original dataset
Ruping [15]		√	√
Cauwenberghs and Poggio [16]		√	√
Fei-Fei et al. [17]	√		√
Tommasi et al. [18]	√		√
Kuzborskij et al.[19]	√		
Engelbrecht and Cloete [20]		√	
Zhang [21]		√	
Ours	√	√	√

Table 2: Experimental results on CK++.

s/m	0.32	0.34	0.36	0.38	0.40	0.42	0.44	0.46	0.48
20	95.89	96.29	96.15	96.27	95.11	93.81	91.28	88.23	77.00
25	95.92	96.18	96.71	96.53	95.35	92.95	91.19	88.37	78.47
30	96.20	96.75	95.45	96.81	95.21	93.08	92.01	87.26	77.38
35	95.78	96.04	95.56	96.79	94.87	92.99	91.78	87.82	76.33
40	95.71	96.21	95.37	96.42	95.12	93.05	90.26	87.11	76.65

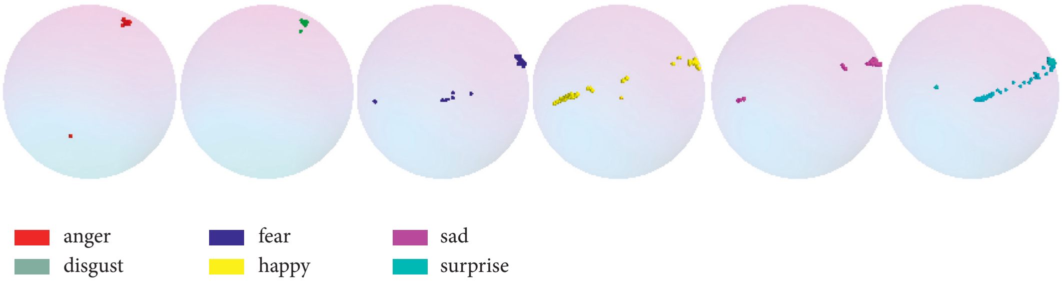

Figure 3: Distribution of extracted features on the sphere for partial sample extraction of CK+ dataset.

4. Experimental Process

We used ExpW as the training set, CK+ as the test set, resnet50 as the backbone network, and softmax as the activation function to train a set of classification models as a control. The optimization method is SGD with a learning rate of 0.001, regularization is used with a regularization rate of 0.0004, and 100 iterations are trained.

In the second experiment, we use ResNet50 as the backbone, remove the last fully connected layer, add a fully connected layer with 512 nodes, and then, add an LMCL layer with 7 output nodes, use ExpW as the training set, CK+ as the test set, and the regularization parameter of the backbone is 0.0004. The weight regularization parameter of the LMCL layer is 0.0005. A grid search was performed for the parameters s and m, with s ranging from 20 to 40 in intervals of 5 and m ranging from 0.3 to 0.5 in intervals of 0.2. Each set of parameters was trained using ExpW for 100 iterations.

We validate the model in the second set of experiments using a composite expression dataset from the raf-db dataset, which includes 11 composite expressions based on 6 basic expression combinations. We randomly select 20 expression pictures from each type of compound expression, use the model to reason 10 of them, obtain feature variables and normalize them, and calculate their mean values. Moreover, a weight category center is added. For the remaining 10 sheets in each category, a total of 110 sheets, after using the model to extract features, the cosine value is calculated with the category center, and the maximum value exceeding the threshold of 0.6 is the corresponding classification.

Table 3: New category center validation set accuracy.

Classification	Validation set accuracy (%)
Happily surprised	70
Happily disgusted	90
Sadly fearful	100
Sadly angry	100
Sadly surprised	80
Sadly disgusted	80
Fearfully angry	80
Fearfully surprised	100
Angrily surprised	100
Angrily disgusted	90
Disgustedly surprised	80
Average	88.18

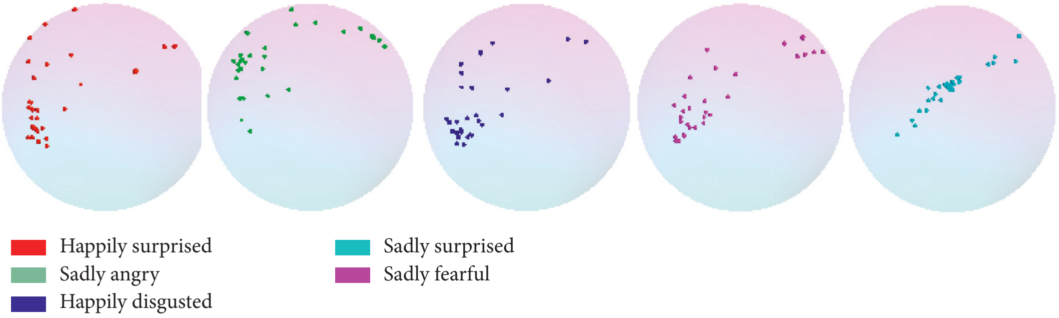

Figure 4: Distribution of features extracted from raf-db custom dataset on the spherical surface.

5. Experimental Results

5.1. Accuracy on CK+

(a) he accuracy of the model on CK+ obtained by standard original softmax loss function training is 97.10%.

(b) Gridding search of parameters m and s based on LMCL training.

As shown in Table 2, it is found that the effect is the best when m is 0.35 and s is 30. When m is greater than 0.42, the accuracy will be greatly reduced. At the same time, s has little effect on the accuracy.

The method of experiment 2 is used, but it is different from adding a fully connected layer with 512 nodes. In order to visualize, a fully connected layer with three nodes is added. The last three expression pictures from each sequence in CK+ dataset are taken to obtain the corresponding three-dimensional feature variables, which are printed on the sphere, as in Figure 3.

It is observed that the feature vectors of the same species achieve class cohesion, there is a clear gap between different species, and the classes are separated by a sufficient distance.

5.2. Calculating a Category Center by Reason a Plurality of Pictures of That User-Defined Category, and Classifying all the Expressions by Using the New Category Center. As shown in Table 3, in the case of using 11 samples to verify, the accuracy of 11 compound expressions reaches 88.18% under 110 verification samples, which can achieve good accuracy without training.

Visualize the distribution of custom categories on the sphere: take 30 samples from each category in the composite expression subset of the raf-db dataset, infer, obtain features on the trained model, normalize them, obtain their features, and print them on the sphere. The distribution of the eigenvectors is observed on the sphere as shown in Figure 4. It is observed that the self-classified data reach the basic class cohesion on the sphere.

6. Conclusion

We reanalyze the geometric significance of the fully connected layer by modifying it, modulo the inputs and weights, and use a small number of samples to allow the model to gain recognition of new categories for scenarios that require customization of new categories. While maintaining the classification accuracy of the original training categories, the newly added categories are also recognized with good accuracy. Compared with the traditional approach of retraining the model to recognize new categories, our approach avoids the cost of expensive data collection and labeling and does not require expensive computational resources and time to retrain the model. However, our proposed method is only applicable when there is correlation between categories and the category features' vector can be added or subtracted. For example, in the domain of expression recognition, the "surprise" expression can be obtained by summing the features of "amazing" and "happy" to get the average value. Therefore, when using this method, it is necessary to evaluate the relationship between categories to determine whether this method can be applied.

Data Availability

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was partially supported by Shenzhen Education Science "14TH FIVE-YEAR PLAN"2021, Research on online learning emotion analysis and intelligent tutoring based on collaborative perception of multimodal education data, key technology research, and innovative application demonstration of intelligent education (2019KZDZX1048), Guangdong Key Laboratory of Big Data Intelligence for Vocational Education (2019GKSYS001), and Shenzhen Vocational Education Research Center Jointly Established by the Ministry and the Province (6022240004Q).

References

[1] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, no. 2, 2012.

[2] X. Li and W. Wang, "Learning discriminative features via weights-biased softmax loss," *Pattern Recognition*, vol. 107, Article ID 107405, 2020.

[3] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proceedings of the European Conference on Computer Vision*, Springer, Amsterdam, The Netherlands, 11 October 2016.

[4] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks," *JMLR.org*, vol. 48, 2016.

[5] H. Wang, Y. Wang, Z. Zhou et al., "CosFace: large Margin cosine loss for deep face recognition," in *Proceedings of the IEEE*, Salt Lake City, UT, USA, 18 June 2018.

[6] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *Journal of Personality and Social Psychology*, vol. 17, no. 2, pp. 124–129, 1971.

[7] X. Liu, B. Kumar, J. You, and p. Jia, "Adaptive deep metric learning for identity-aware facial expression recognition," in *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition Workshops*, IEEE, Honolulu, HI, USA, 21 July 2017.

[8] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, IEEE, Santiago, Chile, 7 December 2015.

[9] P. Khorrami, T. L. Paine, and T. S. Huang, *Do deep neural networks learn facial action units when doing expression recognition*, IEEE, New Jersey, 2015.

[10] H. Ding, S. K. Zhou, and R. Chellappa, *FaceNet2ExpNet: regularizing a deep face recognition net for expression recognition*, IEEE, New Jersey, 2016.

[11] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa, "An all-in-one convolutional neural network for face analysis," in *Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, IEEE, Washington, DC, USA, 30 May 2017.

[12] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): a complete dataset for action unit and emotion-specified expression," in *Proceedings of the Computer Vision & Pattern Recognition Workshops*, IEEE, San Francisco, CA, USA, 13 June 2010.

[13] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "From facial expression recognition to interpersonal relation prediction," *International Journal of Computer Vision*, vol. 126, no. 5, pp. 550–569, 2018.

[14] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Transactions on Image Processing*, vol. 28, p. 1, 2018.

[15] S. Ruping, "Incremental learning with support vector machines," in *Proceedings of the 2001 IEEE International Conference on Data Mining*, pp. 641-642, IEEE, San Jose, CA, USA, 29 November 2001.

[16] G. Cauwenberghs and T. Poggio, "Incremental and decremental support vector machine learning," *Advances in Neural Information Processing Systems*, vol. 13, 2000.

[17] L. Li Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 594–611, 2006.

[18] T. Tommasi, F. Orabona, and B. Caputo, "Learning categories from few examples with multi model knowledge transfer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 5, pp. 928–941, 2014.

[19] I. Kuzborskij, F. Orabona, and B. Caputo, "From N to N+1: multiclass transfer incremental learning," in *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3358–3365, IEEE, Portland, OR, USA, 23 June 2013.

[20] A. P. Engelbrecht and I. Cloete, "Incremental learning using sensitivity analysis," in *Proceedings of the IJCNN'99. International Joint Conference on Neural Networks*, no. 2, pp. 1350–1355, IEEE, Washington, DC, USA, 10 July 1999.

[21] B. T. Zhang, "An incremental learning algorithm that optimizes network size and sample size in one trial," in *Proceedings of the 1994 IEEE International Conference on Neural Networks (ICNN'94)*, no. 1, pp. 215–220, IEEE, Orlando, FL, USA, 28 June 1994.