2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

Batch Normalization Masked Sparse Autoencoder for Robotic Grasping Detection

Zhenzhou Shao[1], Ying Qu[2*], Guangli Ren[3], Guohui Wang[1], Yong Guan[1], Zhiping Shi[1], Jindong Tan[2]

***Abstract*— To improve the accuracy of the grasping detection, this paper proposes a novel detector with batch normalization masked evaluation model. It is designed with a two-layer sparse autoencoder, and a Batch Normalization based mask is incorporated into the second layer of the model to effectively reduce the features with weak correlation. The extracted features from such model are more distinctive, which guarantees the higher accuracy of the grasping detection. Extensive experiments show that the proposed evaluation model outperforms the state-of-the-art, and the recognition accuracy can reach 95.51% for robotic grasping detection.**

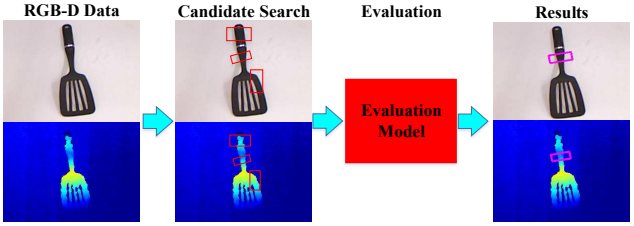


Fig. 1: Implementation of the robotic grasping detection.

I. INTRODUCTION

Robotic grasping has attracted a lot of attention in the field of intelligent robot and contributed significantly in a wide range of application domains, including home service [1], industrial production [2], space exploration [3], etc. One of the important prerequisite of robotic grasping is the detection of a proper grasping position, which is referred to as the robotic grasping detection, as illustrated in Fig. 1. Generally, given RGB-D data, grasping position detection can be converted to a search-evaluation problem using computer vision technique [4], [5]. That is, first, the candidates of grasping positions are selected given an image of the target object, then the optimal position is chosen using the evaluation model. Although grasping position detection has been intensively studied, it remains a challenging task due to the diversity size, material, and the poses of grasping targets [6].

In the past few decades, numerous approaches have been developed to detect the grasping position. Conventional methods solve this problem by introducing the force closure constraint [7], [8], or employing the physical analysis techniques, such as caging [9], [10], [11], grasp wrench space analysis [11], and object wrench space analysis [12]. However, the success of above methods greatly depends on the accurate perception of objects' visual features, *e.g.*, the shape and position information of the objects. And they could not find the optimal grasping position of the objects that do not belong to the training data. There have been several attempts to solve the grasping position detection problem based on the features extracted by learning based approaches. Through hand-designed features, a multi-step method is proposed by Jiang *et al.* [4] to learn the optimal grasping positions of novel objects. However, the pipeline of this approach has to be hand-coded based on the specific features of the new tasks.

Recently, deep learning, which revolutionized many fields of studies, has also been applied to evaluate the grasping positions by learning the characteristics of multi-modalities including both color and depth images [13], [1], [14]. Moreover, in order to achieve better performance, the multi-modality feature fusion [15], [16] is employed to rearrange features into a one-dimensional vector. However, due to the premature fusion of features, the network is prone to be over-fitting.

To address above problems, we propose a batch normalization masked sparse autoencoder (SAE) for robotic grasping detection given RGB-D images. The features are extracted using SAE in an unsupervised way, then the supervised learning is employed to train the evaluation model. The main contribution of this paper is that a novel evaluation model based on the mask defined by batch normalization is proposed to evaluate the reliability of the grasping positions. The model consists of two SAE hidden layers. In the first hidden layer, SAE is employed to fuse the features extracted from regions of grasping candidates. To remove redundant features with weak correlation and prevent overfitting, the features are filtered by a mask matrix generated with Batch Normalization (BN) [17]. Thus, sparse and effective features with distinct characteristics are extracted and fed into the following classification layers to find reliable grasping positions. This layer is referred to as BN-SAE for abbreviation. We evaluate the proposed method on a challenging dataset. Experimental results show that our method outperforms the state-of-the-art on both the accuracy and efficiency of the grasping position detection.

*Corresponding author
[1]Zhenzhou Shao, Guohui Wang, Yong Guan and Zhiping Shi are with the College of Information Engineering, Beijing Advanced Innovation Center for Imaging Technology and Beijing Key Laboratory of Light Industrial Robot and Safety Verification, Capital Normal University, Beijing, 100048, China. {zshao,ghwang, shizp, guanyong}@cnu.edu.cn
[2]Ying Qu and Jindong Tan are with the Engineering College, The University of Tennessee, Knoxville, TN, 37996, USA. {yqu3,tan}@utk.edu
[3]Guangli Song is with the State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China. renguangli2018@ia.ac.cn

II. PROBLEM FORMULATION

When a robot performs a grasping task, a series of grasping rectangles representing the grasping position candidates are generated by the sliding window search given RGB-D images. We use a pre-trained deep neural network as an evaluation model to determine the reliability of these grasping rectangles, according to the features of the color, depth and surface normal vector extracted from each grasping position.

In this paper, the grasping rectangle $G^{(i)}$ is employed to represent the grasping position, the features of $G^{(i)}$ is represented by $X(i)$, during the evaluation process, the reliability of each grasping position is denoted by $y^{(i)} \in \{0, 1\}$, where 1 represents positive position, otherwise negative one. Therefore, the evaluation of grasping positions can be transformed into the problem of solving the probability model defined as

$$\begin{cases} \hat{y}_i^{(1)} = P(y^{(i)} = 1|X(i), W) \\ \hat{y}_i^{(0)} = P(y^{(i)} = 0|X(i), W) \\ y^{(i)} = \max(\hat{y}_i^{(0)}, \hat{y}_i^{(1)}), \end{cases} \tag{1}$$

where W denotes the weights of evaluation model, for an L-layer evaluation network, the relationship between the input and output of the network is:

$$\begin{cases} h_j^{(1)} = sigm(\sum_{m=1}^{K^0} W_{m,j}^{(1)} X_m(i)) \\ h_j^{(k)} = sigm(\sum_{m=1}^{K^{(k-1)}} W_{m,j}^{(k)} h_m^{(k-1)}) \\ P(\hat{y}^{(i)}|X(i), W) = sigm(\sum_{m=1}^{K^{(L-1)}} W_{m,j}^{(L)} h_m^{(L-1)}), \end{cases} \tag{2}$$

where $sigm(x) = 1/(1 + \exp(-x))$, $K^{(k-1)}$ denotes the number of neurons in layer $k-1, k = 1, 2, \ldots, L$, the weight vector $W = (W^1, W^2, \ldots, W^{L-1})$ is learned in an unsupervised way.

The process of detecting the optimal rectangle G^* can be described as follows: In the grasp space $GSpace$ of the target, the grasping rectangle satisfying (3) is used as the optimal grasping rectangle G^*.

$$G^* = arg \max_{G^{(i)}} P(\hat{y}^{(i)} = 1|X(i), W). \tag{3}$$

III. EVALUATION MODEL BASED ON BATCH NORMALIZATION MASKED SPARSE AUTOENCODER

In this section, a novel evaluation model based on batch normalization mask is proposed. The representative features are learned in a unsupervised manner by employing the SAE structure. In order to reduce the computational cost and guarantee the accuracy of grasping detection, Batch Normalization is introduced in SAE to further improve the sparsity of features.

A. BN-SAE: Batch Normalization Masked Sparse Autoencoder

Generally, a L_1 regularization is applied on a traditional sparse autoencoder to reduce the complexity of the model. During the training procedure, the desired features $a = \{a_1, a_2, \ldots, a_n\}$ can be extracted by minimizing the reconstruction error (4) given an input vector $X = \{x_1, x_2, \ldots, x_n\}$.

The objective function is defined as:

$$\min_{a,W,b} \|Wa + b - X\|^2 + \lambda \Sigma_j |a_j|, \tag{4}$$

where W represents the weight matrix between layers, and b is the bias matrix. λ balances the trade-off between the penalty constraint and the reconstruction error. We can adjust the parameter λ to increase the sparsity of the network. However, the learned features tend to lose the correlation with respect to the input data, because most of the features are close to zero due to the large L_1 regularization. In this paper, we propose a batch normalization (BN) [18] based sparse autoencoder (BN-SAE) to address this problem. Instead of using L_1 penalty, we use a mask matrix generated by batch normalization to filter the learned features. The BN-based mask matrix enforces the learned features by the network to be sparse, and retains the correlation between the representation layer and the input data. In addition, it accelerates the training process of the network.

The proposed BN-SAE structure is shown in Fig. 2. During the training procedure, the hidden layer is normalized by their mean μ_x and stranded deviation σ_x as described in (5), (6) and (7). Then the batch normalization matrix z_i is generated by (8). Two parameters γ and β are adjusted automatically while minimizing the reconstruction error. Finally, the mask matrix M_{mask} is generated by (9).

$$\mu_x = \frac{1}{m} \sum_{i=1}^{m} x_i \tag{5}$$

$$\sigma_x^2 = \frac{1}{m} \sum_{i=1}^{m} (x_i - \mu_x)^2 \tag{6}$$

$$\hat{x}_i = \frac{x_i - \mu_x}{\sqrt{\sigma_x^2 + \epsilon}} \tag{7}$$

$$z_i = \gamma \hat{x}_i \equiv BN(x_i, \gamma, \beta) \tag{8}$$

$$M_{mask} = sigm(z_i) \tag{9}$$

As shown in (10), the mask matrix is applied to further improve the sparsity and reduce the redundant features that have less contribution on the results. To better demonstrate

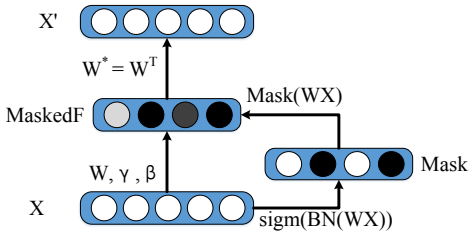
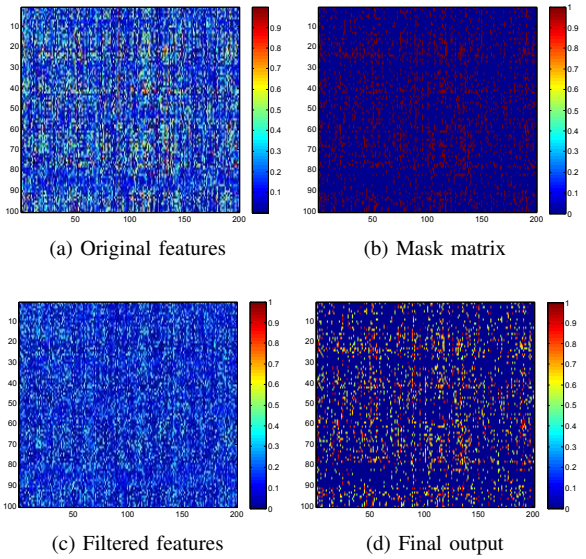


Fig. 2: Structure of BN-SAE.



(a) Original features (b) Mask matrix

(c) Filtered features (d) Final output

Fig. 3: Filtered features by the BN-based mask matrix.

the effects of the BN-SAE mask matrix that filters the features, we use a toy example in the dataset [19] to visualize and compare the features before and after the BN-SAE processing. The output features of the first hidden layer are used as the input of the BN-SAE layer. The results are shown in Fig. 3.

$$F_{masked} = sigm(M_{mask} * WX) \tag{10}$$

We can observe that the values of features filtered by the BN-based mask are mostly small. That is, such features have less contribution on the evaluation results. From Fig. 3d, we can see that the amplitude variation of the remaining features after filtering is more noticeable than the features without the filtering procedure, as shown in Fig. 3a. In other words, after the mask matrix procedure, the features are more vivid and easily to distinguish. More discussions are demonstrated in Sec IV-B.

B. Proposed Evaluation Model with BN-SAE

The evaluation model is a deep learning network based on two SAE based layers. The weights W_1 and W_2 of feature extraction part are obtained by the unsupervised learning, and the weights W_3 for the classification layer are studied by supervised learning.

The structure of proposed evaluation model for grasping detection is shown in Fig. 4. In the network design, the first hidden layer is constructed based on a sparse autoencoder with multi-modal regularization L_W in [1]. It is used to fuse the multi-modal features of the input X. And the weights of SAE are learned to initialize W_1 in the network. X' is the reconstructed result with respect to the first hidden layer. Then the proposed BN-SAE is used as the second hidden layer, where the mask matrix of BN-SAE M_{mask} is used to encourage the sparsity of the features. The weight W_2 in the BN-SAE is taken as the initial weights of the second hidden

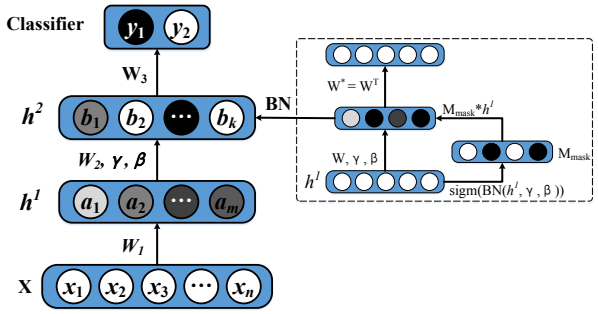


Fig. 4: Structure of the evaluation model.

Algorithm 1 The training procedure of the proposed evaluation model for grasping position with BN-SAE.

Input: A set of the grasping positions containing multi-modality features X with label vector y, the initialization of the weights $W_{1,0}, W_{2,0}, W_{3,0}$, the initialization of the parameters γ_0 and β_0 for batch normalization.

1: Train W_1 based on SAE with multi-modal regularization: $(W_1, X') \leftarrow Train(SAE, X, L_W)$.
2: Calculate the output h^1 of SAE based layer: $h^1 \leftarrow sigm(W_1 X)$.
3: Calculate the initial mask matrix M_{mask}: $M_{mask} \leftarrow sigm(BN(h^1, \gamma_0, \beta_0))$.
4: Train W_2 based on BN-SAE: $(W_2, M_{mask}, \gamma, \beta) \leftarrow Train(\text{BN-SAE}, F_1, M_{mask})$.
5: Calculate the output h^2 of BN-SAE based layer: $h^2 \leftarrow sigm(M_{mask} * (W_2 * h^1))$.
6: Train W_3 in the supervised manner using

$$\begin{cases} \hat{y} \leftarrow Softmax(W_3 * h^2) \\ W_3 = \underset{W_3}{\operatorname{argmin}} \ (y - \hat{y} | W_3). \end{cases} \tag{12}$$

Output: $W_1, W_2, W_3, \gamma, \beta$.

layer. And the parameters γ and β in BN are introduced as the parameters of the final network. For grasping position evaluation, the last layer of the network adopts $Softmax$ as the classification layer to perform the supervised training.

According to (11) and the back propagation mechanism, the network weight W_3 is updated during the training procedure. The grasping position evaluation is summarized in Algorithm 1.

$$P(y = 1 | x; W_3) = sigm(W_3 * h^2) \tag{11}$$

Since the mask matrix filtering mechanism in BN-SAE enforces the network to learn highly-sparse features, we only adopt it in the second hidden layer. If the original input data is processed and filtered twice by the mask matrix, it may force the network to drop out useful information, which reduces the accuracy of the grasping position detection. More details are discussed in Section IV-C.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

The proposed method has been carefully evaluated in two sets of experiments. First, the proposed evaluation model is evaluated compared with five state-of-the-art evaluation methods for grasping position detection. Second, BN-SAE layer is evaluated by investigating the distribution of output features and testing performance.

All experiments are implemented using the benchmarked grasping dataset from Cornell University [19]. The raw dataset contains 885 groups of images of 240 objects including both the color images and point clouds. There are totally 7908 grasping rectangles, containing features and ground truth labels. We randomly split the dataset into training data consisting of 6326 grasping rectangles, and testing data with 1582 grasping rectangles. The seven channels of features are chosen in the experiments, including YUV, surface normal vector and depth measurements. In our experiments, each grasping rectangle has 24×24 pixels and each pixel contains 7 features, thus we set the number of the nodes in the input layer as 4032, *i.e.*, $7 \times 24 \times 24$.

In this section, the proposed evaluation model is compared with five state-of-the-art methods for grasping position, including Chance, Jiang *et al.* [4], Jiang *et al.* + FPFH [1], two-layer sparse AE with L_1 [1] and two-layer sparse AE with group regularization [1]. Similar to [1], we compare our recognition results in the Cornell grasping dataset with the features from above methods, then a linear SVM is adapted for classification. In addition, we also perform comprehensive evaluation of the proposed BN-SAE model with respect to the accuracy, sparsity and efficiency. The intersection-over-union metric is introduced to estimate the grasping results, defined by $IoU = Area(G \bigcap G^*)/Area(G \bigcup G^*)$. When $IoU > 25\%$, the result is considered acceptable.

The proposed network has two hidden layers consisting of SAE and BN-SAE, each of which has 200 hidden units. See detailed network design in Section III-A. The last layer outputs the evaluation score of input grasping rectangle, which is compared with the ground truth, *i.e.*, 1 indicates the positive position and 0 for the negative one.

A. Verification of evaluation model

The most important measurement for the evaluation performance is the accuracy, as it determines whether the final grasping task would succeed or fail. We apply the proposed evaluation model to all objects in the testing data, and the average accuracy of different methods is shown in Table I. The chance performance is obtained by randomly choosing a grasping position and assigning a random score to determine whether the rectangle is graspable or not, which gives a baseline with the accuracy of 50%. The traditional approach [4] that selects features manually, is able to increase the accuracy rate to 84.7%. If the features are selected from both the approach [4] and the Fast Point Feature Histogram (FPFH) [1], the accuracy is increased by 4.9%, but it is still below the deep learning based approaches like SAE or the proposed method. The SAE based approaches, either with L_1 regularization or structured regularization, outperform traditional approaches to a large extend. The proposed method with the BN-SAE structure is able to outperform other state-of-the-art approaches and achieve an accuracy of 95.51%. That is because the network with proposed BN-SAE is capable of reducing redundant features and keeping significant features that contributes to the performance gain. In this way, the detection accuracy is largely improved. More details about the BN-SAE are discussed in Sec IV-B and Sec IV-C.

TABLE I: Recognition results for Cornell grasping dataset.

	Methods	Accuracy
1	Chance [5]	50 %
2	Jiang *et al.* [4]	84.7 %
3	Jiang *et al.* + FPFH [1]	89.6 %
4	Two-layer SAE, L_1 [1]	93.7 %
5	Two-layer SAE, reg. [1]	93.7 %
6	Proposed SAE+BN-SAE	**95.51** %

TABLE II: IoU using different evaluation methods.

Object \ Method	Two-layer SAE	SAE+BN-SAE
Banana	31.24%	**49.07**%
Shovel	26.88%	**58.46**%
Shoe	12.89%	**50.52**%
Brush	25.42%	**33.81**%
Bottle	48.32%	**63.72**%
Bulb	51.78%	**53.65**%

To further demonstrate the effectiveness of the proposed evaluation model for grasping position detection, we perform a grasping detection experiment on individual objects using the proposed model with the pyramid based search in our previous work [21]. In this experiment, we only compare our method (SAE + BN-SAE) with the one proposed by Lenz *et al.* [1] (Two-layer SAE) that has better performance. Fig. 5 shows the optimal grasping positions by applying both methods to commonly benchmarked objects with different physical properties. We observe that the proposed method is able to consistently detect the optimal positions of different types of objects.

For qualitative comparison, the IoU values of the optimal grasping positions are calculated, as shown in Table II. Note that, the located position is considered to be a good grasping position, when IoU is higher than 25%. The IoU values achieved from the proposed SAE + BN-SAE are always higher than 33%. This indicates the proposed approach can always find the optimal position. But the values obtained from the Two-layer SAE are relatively low, and some of the values are even less than 25%, which indicates a negative grasping position. The IoU values visualized in Fig. 6 demonstrate the superiority of the proposed method.

B. Effects of the BN-SAE

In the network design, the number of the input nodes equals to the number of pixels times the number of features in a image patch. Generally, large input nodes make the evaluation model more complex and may introduce overfitting. To prevent overfitting, it is important to encourage

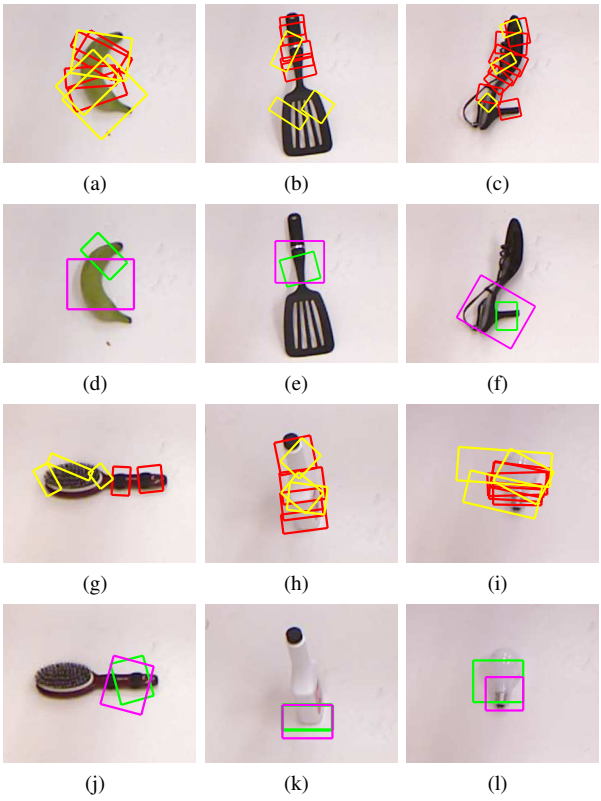


Fig. 5: Optimal grasping positions located by the proposed approach and Lenz *et al.* [1]. (a)-(c) and (g)-(i) show the ground truth given in the dataset. The red and yellow rectangles denote the positive and negative grasping positions, respectively. (d)-(f) and (j)-(l) show the optimal positions detected by different methods. The positions detected by the proposed SAE + BN-SAE model are drawn in green, while the ones detected using two-layer SAE are shown in magenta.

the representations to be sparse. The sparsity constraint also serves to reduce the complexity of the model as well as increase the detection accuracy. We measure the sparsity of the network by the distribution of active values in the hidden layer. As shown in Fig. 7, the horizontal coordinate indicates the activation values of neurons, and the vertical coordinate represents the number of activated neurons in the hidden layer. Note that the total number of neurons are $100 \times 200 = 20000$.

The results indicate that the proposed activation values in the BN-SAE layer is sparser than that of the traditional SAE layer. That is because the mask mechanism preserves high activation values and suppresses low ones. In this way, it encourages the features to be distinctive which improves the accuracy of grasping position detection.

The difference between the training and testing accuracy indicates the ability of networks to handle the over-fitting. As shown in Fig. 8, the accuracy of SAE + BN-SAE can achieve more than 50% in a few iterations, and the learning curve is always higher than that of the two-layer SAE at the same time. In the testing process, the proposed method

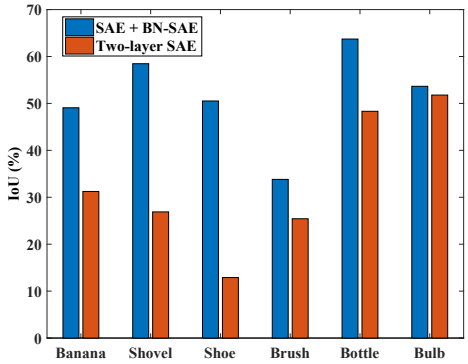
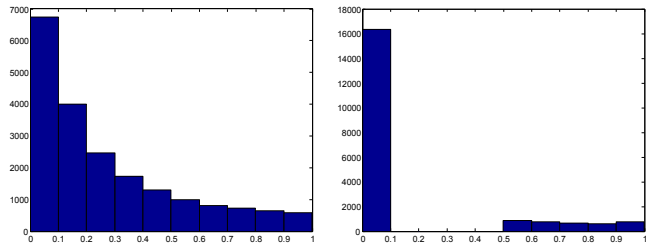


Fig. 6: IoU comparison between the proposed SAE + BN-SAE and the two-layer SAE.



(a) Features extracted by SAE (b) Features extracted by BN-SAE

Fig. 7: The distribution of active neurons in the hidden layer.

is also superior to the two-layer SAE based algorithm. This indicates that our algorithm is more powerful than the two-layer SAE based approach in terms of over-fitting handling.

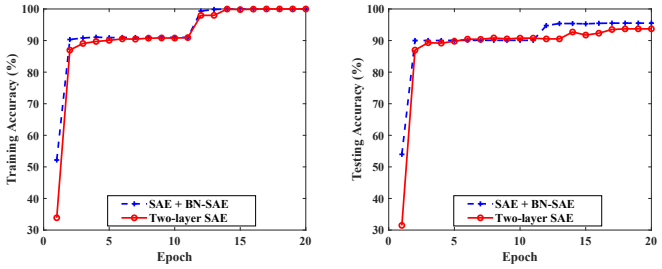


Fig. 8: Training accuracy and testing accuracy. The left figure shows the training accuracy and the right one demonstrates the testing accuracy.

C. Discussion

In the network design, we use one BN-SAE layer concatenated with SAE layer to build the evaluation model. In this section, we discuss and analyze the reason of such design. Assume that all the hidden layers are constructed with the proposed BN-SAE layer. Table III shows the training time and the testing accuracy. The training process with two-layer BN-SAE model is very efficient. However, because the BN-SAE layers filter the data twice, the network losses a large amount of grasping information due to the mask mechanism. Thus the evaluation accuracy is dropped by 8.41%.

TABLE III: Grasping accuracy and traning time.

Methods	Accuracy	Training Time (s)
Two BN-SAE	87.10%	215
Proposed SAE+BN-SAE	95.51%	542

TABLE IV: IoU of different network design.

Object \ Method	Two-layer BN-SAE	SAE+BN-SAE
Banana	19.43%	49.07%
Shovel	0	58.46%
Shoe	18.89%	50.52%

The IoU of optimal grasping positions for the first three objects in Fig. 5 is shown in Table IV. Again, to intuitively visualize the gap between both network designs, we show its distribution in Fig. 9. We can observe that the IoU of the shovel is zero, meaning that the optimal position acquired from two-layer BN-SAE is completely unacceptable. In addition, the optimal positions achieved from the two BN-SAE based evaluation model are all unacceptable. Therefore, in our network design, only one BN-SAE layer is adopted as the second hidden layer to train the evaluation model.

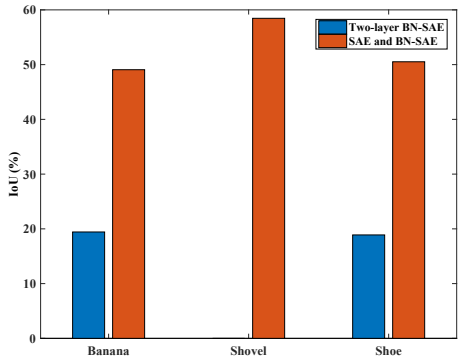


Fig. 9: IoU comparison between two-layer BN-SAE and SAE +BN-SAE.

V. CONCLUSION

This paper proposes a batch normalization masked sparse autoencoder for robotic grasping detection. The Batch Normalization is introduced as a feature mask in the second hidden layer in the evaluation model to reduce the redundancy of feature representations. It not only improves the accuracy of the grasping detection, but also accelerates the training procedure, as the computational load for distinct features is reduced. Experimental results demonstrate the superiority of the proposed approach compared to state-of-the-art.

ACKNOWLEDGMENT

This work was supported by National Key R & D Program of China (2019YFB1309900), National Natural Science Foundation of China (61702348, 61772351), Beijing Nova Program of Science and Technology (Z191100001119075), the National Technology Innovation Special Zone (19-163-11-ZT-001-005-06) and Academy for Multidisciplinary Studies, Capital Normal University(19530012005).

REFERENCES

[1] Ian Lenz, Honglak Lee, and Ashutosh Saxena. Deep learning for detecting robotic grasps. *International Journal of Robotics Research*, 34(4-5):705–724, 2015.

[2] A. Kramberger, A. Wolniakowski, M. H. Rasmussen, M. Munih, A. Ude, and C. Schlette. Automatic fingertip exchange system for robotic grasping in flexible production processes. In *2019 IEEE 15th International Conference on Automation Science and Engineering (CASE)*, pages 1664–1669, 2019.

[3] Wenya Wan, Chong Sun, Jianping Yuan, Xianghao Hou, Yufei Guo, Yinong Ou-yang, Qixin Li, Liran Zhao, Hao Shi, and Dawei Han. Adaptive whole-arm grasping approach of tumbling space debris by two coordinated hyper-redundant manipulators. In Haibin Yu, Jinguo Liu, Lianqing Liu, Zhaojie Ju, Yuwang Liu, and Dalin Zhou, editors, *Intelligent Robotics and Applications*, pages 450–461, 2019.

[4] Yun Jiang, S Moseson, and A Saxena. Efficient grasping from rgbd images: Learning using a new rectangle representation. In *IEEE International Conference on Robotics and Automation*, pages 3304–3311, 2011.

[5] Zhichao Wang, Zhiqi Li, Bin Wang, and Hong Liu. Robot grasp detection using multimodal deep convolutional neural networks. *Advances in Mechanical Engineering*, 8(9):1–12, 2016.

[6] Jeffrey Mahler, Florian T. Pokorny, Brian Hou, Melrose Roderick, Michael Laskey, Mathieu Aubry, Kai Kohlhoff, Torsten Kröger, James J. Kuffner, and Kenneth Y. Goldberg. Dex-net 1.0: A cloud-based network of 3d objects for robust grasp planning using a multi-armed bandit model with correlated rewards. *2016 IEEE International Conference on Robotics and Automation*, pages 1957–1964, 2016.

[7] Noé Alvarado Tovar and Raúl Suárez. Searching force-closure optimal grasps of articulated 2d objects with n links. *IFAC Proceedings Volumes*, 47(3):9334–9340, 2014.

[8] Daniel Kappler, Jeannette Bohg, and Stefan Schaal. Leveraging big data for grasp planning. In *2015 IEEE International Conference on Robotics and Automation*, pages 4304–4311. IEEE, 2015.

[9] Rosen Diankov, Siddhartha S Srinivasa, Dave Ferguson, and James Kuffner. Manipulation planning with caging grasps. In *Humanoid Robots, 2008. Humanoids 2008. 8th IEEE-RAS International Conference on*, pages 285–292. IEEE, 2008.

[10] Alberto Rodriguez, Matthew T Mason, and Steve Ferry. From caging to grasping. *The International Journal of Robotics Research*, 31(7):886–900, 2012.

[11] Máximo A Roa and Raúl Suárez. Grasp quality measures: review and performance. *Autonomous robots*, 38(1):65–88, 2015.

[12] Shuo Liu and Stefano Carpin. A fast algorithm for grasp quality evaluation using the object wrench space. In *2015 IEEE International Conference on Automation Science and Engineering*, pages 558–563. IEEE, 2015.

[13] Lars Berscheid, Pascal Meißner, and Torsten Kröger. Robot learning of shifting objects for grasping in cluttered environments. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 612–618. IEEE, 2019.

[14] U. Asif, M. Bennamoun, and F. A. Sohel. Rgb-d object recognition and grasp detection using hierarchical cascaded forests. *IEEE Transactions on Robotics*, 33(3):547–564, 2017.

[15] Le An, Xiaojing Chen, and Songfan Yang. Multi-graph feature level fusion for person re-identification. *Neurocomputing*, 259:39–45, 2017.

[16] S. Kumra and C. Kanan. Robotic grasp detection using deep convolutional neural networks. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 769–776, Sep. 2017.

[17] 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, 2015.

[18] 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, 2015.

[19] Cornell grasping dataset. `http://pr.cs.cornell.edu/grasping/rect_data/data.php`.

[20] Hakan Karaoguz and Patric Jensfelt. Object detection approach for robot grasp detection. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 4953–4959. IEEE, 2019.

[21] G. Ren, Z. Shao, Y. Guan, Y. Qu, J. Tan, H. Wei, and G. Tong. A fast search algorithm based on image pyramid for robotic grasping. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6520–6525, Sep. 2017.