SUGISAKA MASANORI

Journal of Robotics, Networking and Artificial Life
Vol. 10(1); June (2023), pp. 33–38
ISSN (Online): 2352-6386; ISSN (Print): 2405-9021
https://alife-robotics.org/jrnal.html

Research Article

Predicting the Weight of Grappling Noodle-like Objects using Vision Transformer and Autoencoder

Nattapat Koomklang[1], Prem Gamolped[1], Eiji Hayashi[1], Abbe Mowshowitz[2]
[1]*Department of Mechanical Information Science and Technology, Kyushu Institute of Technology, 680-4 Kawazu, Iizuka, Fukuoka 820-8502, Japan*
[2]*Department of Computer Science, The City College of New York, 160 Convent Avenue, New York, NY 10031, USA*

ARTICLE INFO

Article History
Received 02 December 2022
Accepted 17 August 2023

Keywords
Robotic manipulation
Weight estimation
Noodle-like objects
Vision transformer
Autoencoder
RGB-D encoding
Deep learning
Transformer network

ABSTRACT

This research paper presents a novel approach for accurate weight estimation in robotic manipulation of noodle-like objects. The proposed approach combines vision transformer and autoencoder techniques with action data and RGB-D encoding to enhance the capabilities of robots in manipulating objects with varying weights. A deep-learning neural network is introduced to estimate the grasping action of a robot for picking up noodle-like objects using RGB-D camera input, a 6-finger gripper, and Cartesian movement. The hardware setup and characteristics of the noodle-like objects are described. The study builds upon previous work in RGB-D perception, weight estimation, and deep learning, addressing the limitations of existing methods by incorporating robot actions. The effectiveness of vision transformers, autoencoders, self-supervised deep reinforcement learning, and deep residual learning in robotic manipulation is discussed. The proposed approach leverages the Transformer network to encode sequential and spatial information for weight estimation. Experimental evaluation on a dataset of 20,000 samples collected from real environments demonstrates the effectiveness and accuracy of the proposed approach in grappling noodle-like objects. This research contributes to advancements in robotic manipulation, enabling robots to manipulate objects with varying weights in real-world scenarios.

1. Introduction

Robotic manipulation of noodle-like objects presents a significant challenge due to their complex and deformable nature. Accurate weight estimation is crucial for the successful grasping and manipulation of such objects. This academic journal introduces a novel approach that combines vision transformer and autoencoder techniques with action data and RGB-D encoding to predict the weight of grappling noodle-like objects. By leveraging deep learning and encoding methods, the proposed approach aims to enhance the capabilities of robots in manipulating objects with varying weights, contributing to advancements in robotic manipulation in real-world scenarios.

This research paper presents a deep-learning neural network designed to estimate the grasping action of a robot for picking up noodle-like objects. The estimation process utilizes RGB-D camera input, while the robot employs a 6-finger gripper and Cartesian movement for manipulation. The hardware setup is illustrated in Figure 1. The noodle-like objects used in the experiments are created using rubber bands with dimensions of 8 mm in width and 100 mm in length.

The study builds upon several research studies that have explored various aspects of RGB-D perception, weight estimation, and deep learning techniques in the context of robotic manipulation [1], [2], [3], [4], [5], [14], [15], [16], [17], [18], [19], [20]. While these studies have shown good performance, they do not incorporate robot actions for weight estimation,

Corresponding author's E-mail: koomklang.nattapat417@mail.kyutech.jp URL: http://www.kyutech.ac.jp/

which is a crucial factor in achieving accurate predictions.

In the field of computer vision, vision transformers have emerged as a powerful architecture for various tasks. Notably, Dosovitskiy et al. [6] introduced the vision transformer, which effectively captures long-range dependencies in images using self-attention mechanisms. This architecture has demonstrated its effectiveness in encoding visual information and extracting meaningful features.

Autoencoders [7] have gained significant attention in machine learning and robotics. These neural networks are specifically designed for unsupervised learning tasks such as data compression and feature extraction. By training an autoencoder on a dataset, it learns to encode the input data into a lower-dimensional latent space representation and decode it back to the original input. In the context of robotic manipulation, autoencoders have been successfully applied to encode and process various sensor inputs, including RGB-D data, enabling more effective analysis and prediction.

In the domain of robotic manipulation, Zeng et al. [8] explored self-supervised deep reinforcement learning to learn synergies between pushing and grasping actions, facilitating the manipulation of noodle-like objects. Additionally, He et al. [9] introduced deep residual learning, which enables the training of very deep convolutional neural networks by utilizing residual connections, leading to remarkable performance in image recognition tasks.

Significant contributions have also been made to the broader field of computer vision. For instance, LeCun et al. [10] developed gradient-based learning algorithms that laid the foundation for modern deep learning approaches, particularly in document recognition. Simonyan et al. [11] proposed very deep convolutional networks for large-scale image recognition, achieving state-of-the-art performance on benchmark datasets. Moreover, Krizhevsky et al. [12] revolutionized image classification with the introduction of the AlexNet architecture, leveraging deep convolutional neural networks.

Inspired by the successes of the Transformer network [13] in natural language processing, this research adapts the Transformer architecture to address the weight estimation problem in robotic manipulation. The Transformer's ability to capture long-range dependencies and relationships within the input data makes it well-suited for encoding sequential and spatial information.

By leveraging these advancements and incorporating the proposed approach, this research presents a framework for weight estimation in robotic manipulation, utilizing the encoding of RGB-D data and robot actions. The aim is to provide a robust and accurate weight estimation framework for grappling noodle-like objects.

The research utilizes datasets collected from random actions in real environments with simulated noodle-like objects, comprising a total of 20,000 data samples.

Fig. 1 6-fingers gripper and noodle-like objects

2. Methods

2.1. *Autoencoder*

In this research, an autoencoder is employed to reduce the size of the data. Specifically, the autoencoder is utilized to merge the RGB and Depth data into a single vector representation. The encoding process involves leveraging the Vision Transformer to encode the RGB and Depth data separately. Subsequently, the encoded RGB and Depth data are concatenated into a unified vector representation.

For the decoding stage, the encoding data is passed through separate fully connected networks—one for RGB and another for Depth. These networks facilitate the reconstruction of the RGB and Depth images through a Convolution Transpose Network.

The network architecture depicting the autoencoder structure and its components is illustrated in Figure 2. This framework enables the transformation of the input RGB and Depth data into a unified and compressed

representation, facilitating accurate weight estimation for noodle-like objects.

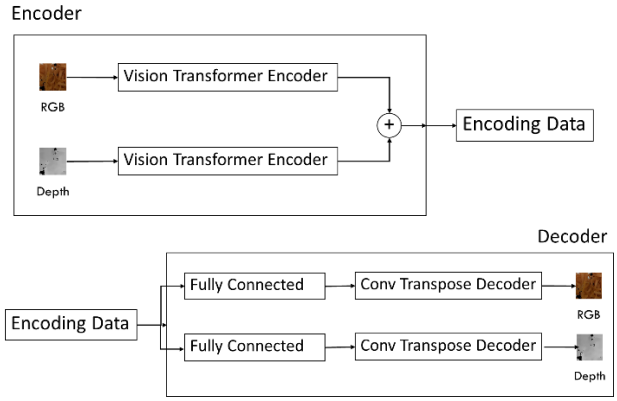


Fig. 2 Encoder and Decoder Network

The autoencoder in this research uses a loss function based on the sum of mean squared errors between the reconstructed RGB and Depth images and their corresponding original images.

$$\text{Loss} = MSELoss_{RGB} + MSELoss_{Depth} \quad (1)$$

In this research, an autoencoder is used to encode the input images. The images are resized from 50x50 pixels to 224x224 pixels and then encoded into a 1024-dimensional vector representation. The encoder employs the Vision Transformer Encoder with the "vit_b_16" architecture. For decoding, a convolution transpose network with a kernel size of 3 is used, along with two convolutional networks to gradually upscale the image size back to 64x64 pixels. The decoder network is shown in Figure 3. The final reconstructed images are cropped to their original size of 50x50 pixels. This autoencoder framework enables the compression of images into a lower-dimensional space while retaining important features for accurate weight estimation. The Vision Transformer Encoder and decoder components work together to encode and decode the images, facilitating subsequent weight estimation based on the encoded RGB-D data and robot actions.

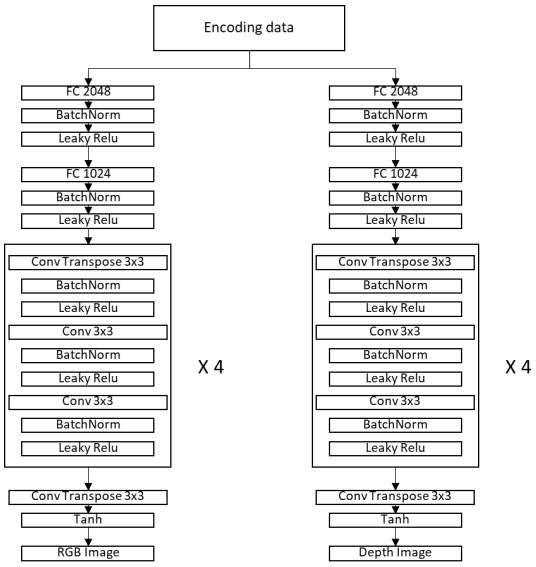


Fig. 3 Decoder Network

2.2. *Transformer Encoder*

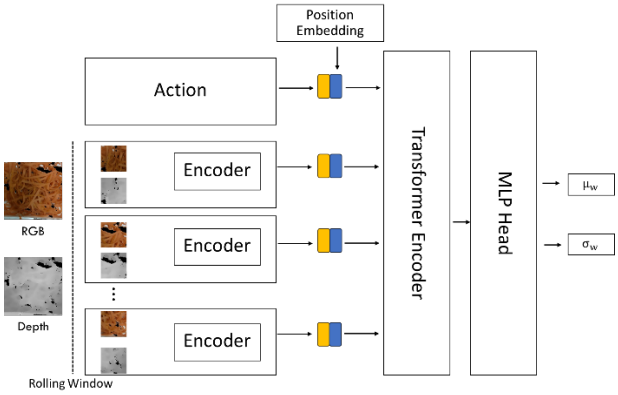


Fig. 4 Weight Estimate Network

The Transformer encoder is a pivotal component of the weight estimation process. Comprising multiple layers, it consists of self-attention mechanisms and feed-forward neural networks. The self-attention mechanism enables the encoder to capture interdependencies between different elements in the input sequence, dynamically assigning varying weights to each element based on its relevance to weight estimation. This mechanism facilitates the modeling of long-range dependencies and the incorporation of global contextual information.

In the weight estimation task, the Transformer encoder is specifically employed. The input to the encoder consists of two layers: one for encoding the action data,

encompassing gripper width and gripper depth, and the other for incorporating a portion of the RGB-D encoding data obtained from the autoencoder. The network is show in Figure 4.

The output of the Transformer encoder is the weight estimation. In this research, weight estimation is performed using a normal distribution, represented by the mean and standard deviation. This approach provides a probabilistic representation of the weight estimation, allowing for a comprehensive understanding of the uncertainty associated with the predictions.

To optimize the weight estimation process, the maximum likelihood estimation of the weight estimation probability is utilized as the loss function. This loss function enables the model to learn the parameters that maximize the likelihood of the observed weight estimation given the input data. By minimizing this loss, the network is trained to generate weight estimations that align with the observed data.

$$\text{Loss} = \log\big(\mathrm{p}(w|x)\big) \qquad (2)$$

By incorporating the Transformer encoder, the proposed approach leverages its ability to capture complex dependencies and incorporate global context information for accurate weight estimation. The utilization of a normal distribution and the maximum likelihood loss function further enhance the reliability and effectiveness of the weight estimation process.

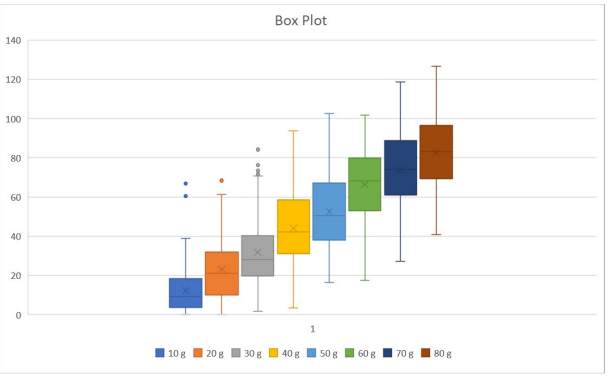


Fig. 5 Picking weight in experimental at 10 – 80 target weight

3. Experiment

In the experiment, the autoencoder component is trained for 10,000 epochs to optimize its encoding capabilities. This extended training duration was chosen based on the convergence behavior observed from the learning curve. Despite the learning curve showing a plateau in updates, the decision to continue training for 10,000 epochs was motivated by the desire to ensure a thorough exploration of the encoding potential and the preservation of vital features in the data. Subsequently, the encoder network harnesses the enriched RGB-D data achieved through this training, which is then employed for weight estimation. This approach aligns with the objective of maximizing the encoder's ability to capture significant information from the input data, enhancing the precision of subsequent weight estimations.

To evaluate the model's performance, a training duration of 2,000 epochs is employed, guided by observations from the learning curve which suggests that further epochs might not significantly enhance updates. During this training, a sampling approach is utilized to expose the network to a diverse range of possible actions. The decision to limit training to 2,000 epochs is aimed at optimizing the trade-off between comprehensive action exploration and computational efficiency. This approach involves varying the gripper width from 30 to 80 mm with a 2.5 mm step size, and the gripper depth from 0.025 to 0.020 m with a 0.025 m step size, along with x and y coordinates selected at 1 cm intervals on a grid, encompassing an extensive array of action scenarios for evaluation.

The weight estimation is performed by selecting the action with the highest likelihood value. For data collection, target weights of 10, 20, 30, 40, 50, 60, 70, and 80 g are set. Approximately 100 data points are collected for each target weight, ensuring a sufficient dataset for evaluation.

By training the network with these settings and performing weight estimation using the maximum likelihood approach, the experiment aims to evaluate the model's ability to accurately estimate the weight of the noodle-like objects. The collected data points across different target weights provide a diverse range of weight scenarios for comprehensive evaluation and analysis.

4. Results

Table 1: The results of grappling test

	10 g	20 g	30 g	40 g	50 g	60 g	70 g	80 g
Mean	12.3	23.1	31.9	43.9	52.6	66.2	73.8	82.5
S. D	11.9	14.7	17.4	18.2	18.5	18.1	19.7	18.0

The results of the experiments are presented in Figure 5 and Table 1, displaying the picking weight for each target weight. The experimental outcomes indicate that the weight estimation tends to align with the specified target

weights. However, it is observed that the distribution of the estimated weights is relatively wide. This variability in the results could be attributed to the sampling process, as it may not encompass a sufficient number of action points to accurately capture the real actions.

Figure 5 provides a visual representation of the estimated picking weights for each target weight, allowing for a quick overview of the weight estimation performance. Table 1 presents the numerical values of the estimated weights corresponding to each target weight, facilitating a detailed analysis of the results.

While the weight estimations generally align with the specified target weights, the relatively high distribution suggests the need for further refinement and improvement in the sampling process. Obtaining a more comprehensive and diverse set of action points would likely enhance the accuracy and precision of the weight estimation.

These results shed light on the current performance of the weight estimation approach and highlight potential areas for future investigation to optimize the sampling strategy and reduce the variability in the estimated weights.
Figure 5 and Table 1 shows the picking weight at each target weight. From the experimental results, it tends to be picked according to the specified weight. But its distribution is relatively high may be due to the sampling for the practical, not enough action point to get the real action.

5. Conclusion

In this paper, we introduced a Transformer encoder as a means to measure the confidence score for evaluating the action of picking noodle-like objects with specified weights. The experimental results demonstrate a tendency for the model to align with the specified weight targets. However, it is worth noting that the distribution of the estimated weights is relatively high.

The utilization of the Transformer encoder in weight estimation showcases its effectiveness in capturing complex dependencies and incorporating global context information. This approach provides valuable insights into the weight estimation process for robotic manipulation tasks.

While the results show promise in aligning with the specified weights, the observed variability in the estimated weights suggests the need for further investigation and refinement. Future research endeavors should focus on refining the model and exploring techniques to reduce the distribution and enhance the precision of the weight estimations.

In conclusion, this paper contributes to the field of weight estimation in robotic manipulation by proposing the use of a Transformer encoder. The presented results provide a foundation for further advancements in accurately evaluating the action of picking noodle-like objects with specified weights.

References

[1] H. Cao, G. Chen, Z. Li, J. Lin and A. Knoll, Lightweight Convolutional Neural Network with Gaussian-based Grasping Representation for Robotic Grasping Detec*tion,* arXiv, 2021.
[2] F.-J. Chu, R. Xu and P. A. Vela, Real-world Multi-object, Multi-grasp Detection, arXiv, 2018.
[3] I. Lenz, H. Lee and A. Saxena, Deep Learning for Detecting Robotic Grasps*,* arXiv, 2013.
[4] D. Liu, X. Tao, L. Yuan, Y. Du and M. Cong, "Robotic Objects Detection and Grasping in Clutter Based on Cascaded Deep Convolutional Neural Network," *IEEE Transactions on Instrumentation and Measurement,* vol. 71, pp. 1-10, 2022.
[5] G. Wu, W. Chen, H. Cheng, W. Zuo, D. Zhang and J. You, "Multi-Object Grasping Detection With Hierarchical Feature Fusion," *IEEE Access,* vol. 7, pp. 43884-43894, 2019.
[6] Dosovitskiy, A., et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations (ICLR), 2021
[7] BANK, Dor; KOENIGSTEIN, Noam; GIRYES, Raja. Autoencoders. arXiv preprint arXiv:2003.05991, 2020.
[8] Zeng, A., et al. (2018). Learning Synergies between Pushing and Grasping with Self-supervised Deep Reinforcement Learning. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), 7337-7344.
[9] He, K., et al. (2016). Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770-778.
[10] LeCun, Y., et al. (1998). Gradient-Based Learning Applied to Document Recognition. Proceedings of the IEEE, 86(11), 2278-2324.
[11] Simonyan, K., et al. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv preprint arXiv:1409.1556.
[12] Krizhevsky, A., et al. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In Advances in Neural Information Processing Systems (NIPS), 1097-1105.
[13] VASWANI, Ashish, et al. Attention is all you need. Advances in neural information processing systems, 2017, 30.
[14] ZENG, Andy, et al. Tossingbot: Learning to throw arbitrary objects with residual physics. IEEE Transactions on Robotics, 2020, 36.4: 1307-1319.
[15] MOUSAVIAN, Arsalan; EPPNER, Clemens; FOX, Dieter. 6-dof graspnet: Variational grasp generation for object manipulation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019. p. 2901-2910.

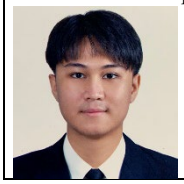
[16] MAHLER, Jeffrey, et al. Learning ambidextrous robot grasping policies. Science Robotics, 2019, 4.26: eaau4984.
[17] H. -Y. Kuo, H. -R. Su, S. -H. Lai and C. -C. Wu, "3D object detection and pose estimation from depth image for robotic bin picking," 2014 IEEE International Conference on Automation Science and Engineering (CASE), New Taipei, Taiwan, 2014, pp. 1264-1269, doi: 10.1109/CoASE.2014.6899489.
[18] SCHILLINGER, Philipp, et al. Model-free Grasping with Multi-Suction Cup Grippers for Robotic Bin Picking. arXiv preprint arXiv:2307.16488, 2023.
[19] H. Cao et al., "Two-Stage Grasping: A New Bin Picking Framework for Small Objects," 2023 IEEE International Conference on Robotics and Automation (ICRA), London, United Kingdom, 2023, pp. 2584-2590, doi: 10.1109/ICRA48891.2023.10160608.
[20] X. Li et al., "A Sim-to-Real Object Recognition and Localization Framework for Industrial Robotic Bin Picking," in IEEE Robotics and Automation Letters, vol. 7, no. 2, pp. 3961-3968, April 2022, doi: 10.1109/LRA.2022.3149026.

Authors Introduction

Mr. Nattapat Koomklang

He received master's degree in engineering in 2018, King Mongkut's University of Technology Ladkrabang in Thailand. He is currently a doctor student at Kyushu Institute of Technology and conducts research at Hayashi Laboratory.

Mr. Prem Gamolped



He received master's degree in engineering in 2021 from Kyushu Institute of Technology. He is currently a doctor student at Kyushu Institute of Technology and conducts research at Hayashi Laboratory.

Prof. Eiji Hayashi

He is a professor in the Department of Intelligent and Control Systems at Kyushu Institute of Technology. He received the Ph.D. (Dr. Eng.) degree from Waseda University in 1996. His research interests include Intelligent mechanics, Mechanical systems and Perceptual information processing. He is a member of The Institute of Electrical and Electronics Engineers (IEEE) and The Japan Society of Mechanical Engineers (JSME).

Prof. Abbe Mowshowitz

He received the Ph.D. degree from University of Michigan in 1967. He has been professor of computer science at the City College of New York and member of the doctoral faculty at the Graduate Center of the City University of New York since 1984. His current research interests lie in two areas are organizational and managerial issues in computing, and network science. In addition to teaching and research, He has acted as consultant on the uses and impacts of information technology (especially computer networks) to a wide range of public and private organizations in North America and Europe.