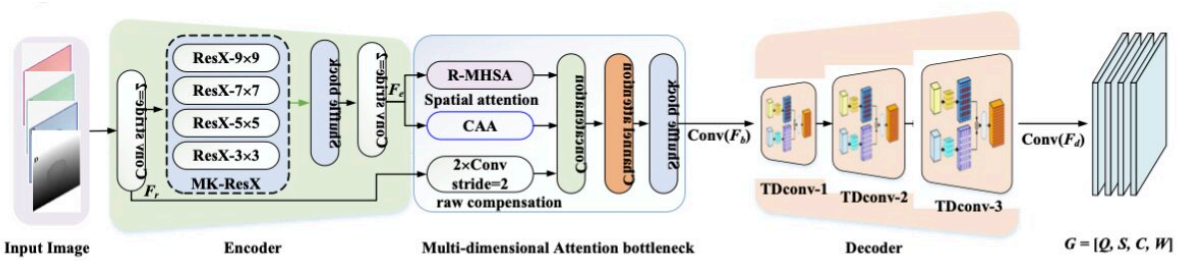
A model that could improve robots' ability to grasp objects

February 6 2023, by Ingrid Fadelli



The framework of the proposed network. Given an RGB-D input image, features are extracted by the encoder. The output feature map F_e of the encoder is further refined through a multi-dimensional attention bottleneck, where the outputs from the residual multi-head self-attention (R-MHSA), cross-amplitude attention (CAA), and raw compensation are concatenated in channel, which is then adjusted by the channel attention and a shuffle block for better feature representation F_b. Followed by a convolution operation, the feature map is fed into the decoder, which adopts three cascaded twin deconvolutions TDconv-1, TDconv-2, and TDconv-3 for grasp prediction G. Credit: Ren et al

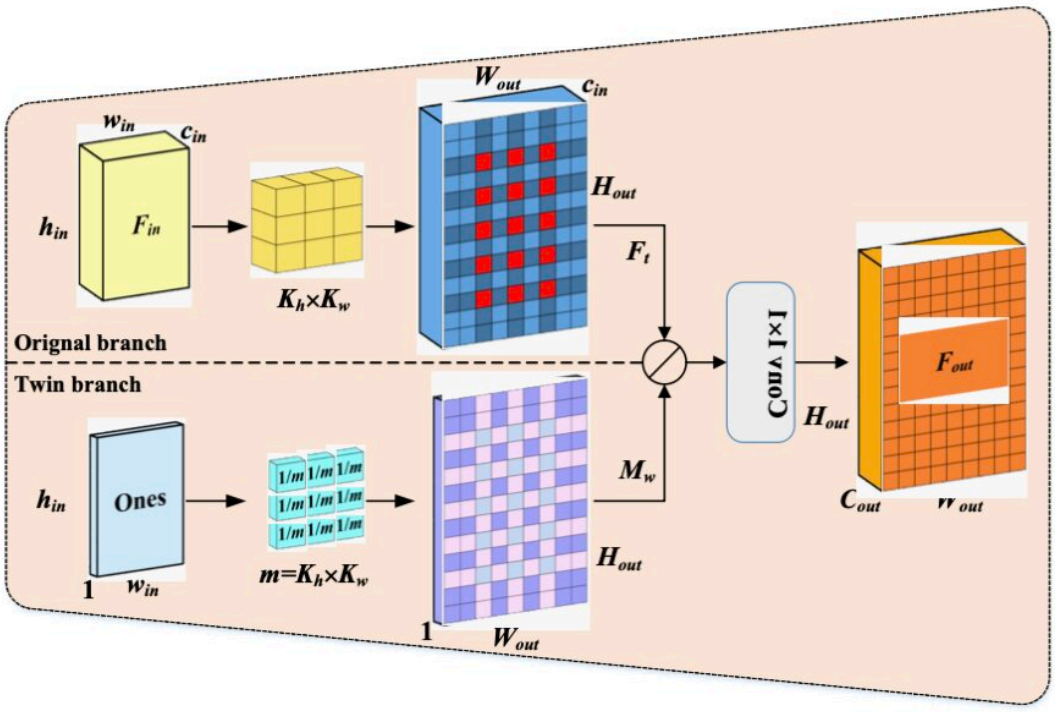
When completing missions and tasks in the real-world, robots should ideally be able to effectively grasp objects of various shapes and compositions. So far, however, most robots can only grasp specific types of objects.

Researchers at the Chinese Academy of Sciences and Peking University

have recently developed a new machine learning model that could help to enhance the grasping abilities of robots. This model, presented in *IEEE Transactions on Circuits and Systems for Video Technology*, is specifically designed to predict grasps for objects in a robot's surroundings, so that they can devise optimal strategies for grasping these objects.

"In real-world applications, such as intelligent manufacturing, human-machine interaction and domestic services, robotic grasping is becoming more and more essential," Junzhi Yu, one of the researchers who carried out the study, told Tech Xplore. "Grasp detection, a critical step of robotic grasping, entails finding the best grasp for a target object. Mainstream encoder-decoder grasp detection solutions are attractive in terms of accuracy and efficiency, yet they are still limited, due to the checkerboard artifacts from uneven overlap of convolution results in the decoder. Moreover, feature representation is often insufficient."

The key objective of the recent work by Yu and his colleagues was develop a model that would overcome the limitations of existing grasp detection frameworks. To do this, they created a pixel-wise grasp detection method based on twin deconvolution and multi-dimensional attention, two established techniques often used for computer vision applications.



Structure of a twin deconvolution. F_in∈R^(c_in×h_in×w_in) and F_out∈R^(C_out ?×H?_out×W_out) denote the input feature map and output feature map, respectively, where c_in,h_in,w_in and C_out,H_out,W_out are the channel number, height, and width corresponding to F_in and F_out. There are two branches in a twin deconvolution: original branch and twin branch, where the former is a standard transposed convolution and the latter is used to calculate the overlap degree corresponding to the original branch for removing checkerboard artifacts. The input of twin branch is a matrix Ones∈R^(1×h_in×w_in) with all the entries 1, whose spatial size is the same as that of the input feature map F_in of the original branch. Moreover, the kernel of the twin branch has the same spatial size as that of the original branch (K_h×K_w) and its all entries are set to 1⁄m, m=K_h×K_w. With the transposed convolution in the twin branch, the overlap degree matrix M_w∈R^(1×H_out×W_out) is computed corresponding to all spatial positions of the output F_t∈R^(c_in×H_out×W_out) from the original branch. Then, an element-wise division operation is performed between each channel of F_t and M_w. Followed by a pointwise convolution Conv1×1, the final output F_out of twin deconvolution is obtained. Credit: Ren et al

Their method was designed to eliminate so-called "checkerboard artifacts," strange checkerboard-like patterns that are often observed in images generated by artificial neural networks. In addition, the researchers strengthened their model's ability to refine specific features in images.

"The proposed pixel-wise grasp detection network is composed of an encoder, a multi-dimensional attention bottleneck, and a twin deconvolution-based decoder," Yu explained. "Given an input image, feature extraction is performed through the encoder and the obtained feature map is further refined through our bottleneck module, which integrates residual multi-head self-attention (R-MHSA), cross-amplitude attention (CAA), and raw compensation to better focus on the regions of interest."

The three components of the team's bottleneck module result in three different outputs that are concatenated in channel and further adjusted to improve the representation of features. The resulting, refined "feature map" is then fed to the model's decoder (i.e., a model that up-samples the feature map into a desirable output). This decoder ultimately predicts the grasps that correspond to the input image, by performing three so-called cascaded twin deconvolutions (processes to up-sample the feature map).

"Through our bottleneck module, the intrinsic relationship between features is mined and features are effectively fine-tuned from the dimensions of spatial and channel," Yu said. "Particularly, the introduction of twin deconvolution provides better up-sampling by adding a twin branch upon original transposed convolution branch. As a result, the challenge of checkerboard artifacts is solved."

The grasp detection on an actual scene. (a) RGB image. (b) Detection results of concerned objects based on Mask R-CNN and background suppression. (c) Grasp detection results. In the experiment scene, the objects with four categories (bottle, banana, apple, and orange) are concerned. Credit: Ren et al

A notable advantage of the method developed by the researchers is its use of twin deconvolutions, through which a twin branch is introduced to the original transposed convolution branch, improving the model's original output. This approach allows the model to remove undesirable checkerboard patterns from outputs.

"It should be noted that the checkerboard artifacts originate from the uneven overlap of convolution results at different positions," Yu said. "Herein, a twin branch is introduced in parallel upon the original transposed convolution branch to measure the uneven overlap. More specifically, the twin branch computes the relative overlap differences among positions and the resultant overlap degree matrix is utilized to re-weight the feature map of original transposed convolution."

In initial tests, the new pixel-wise grasp detection method achieved very promising results, as it was found to smoothen the model's original output and eliminate checkerboard artifacts. It thus achieved a high grasp detection accuracy.

As part of their study, Yu and his colleagues were also able to extend their approach to other task that entail pixel-wise detection. In addition to potentially enhancing the grasping skills of both existing and newly developed robots, their model could thus soon be applied to other computer vision problems.

"In our next works, we plan to combine the proposed method with instance segmentation in actual robot systems for better grasp prediction," Yu added. "For example, instance segmentation can be used to generate valuable information about object profile and position, which is fed into twin deconvolutions of the decoder to further improve the network performance."

More information: Guangli Ren et al, Pixel-wise Grasp Detection via Twin Deconvolution and Multi-Dimensional Attention, *IEEE Transactions on Circuits and Systems for Video Technology* (2023). DOI: 10.1109/TCSVT.2023.3237866