EfficientViT for Retail

Daniel Kim, Jason Li, Joey Zheng
Massachusetts Institute of Technology

Abstract

AI segmentation has found a variety of useful applications across industries. Automotive companies use segmentation to guide self-driving vehicles. Other companies use segmentation for security and surveillance or even virtual reality applications. With that being said, many segmentation models require large computing resources and can only be deployed to devices with significant computational resources. We propose the application of EfficientViT, a lightweight family of vision transformer models that can be run on smaller more limited compute devices, for the particular task of retail segmentation. We use EfficientViT to segment common retail items, allowing them to be classified by a downstream classification model. Finally, we deploy our model to a device with limited computing resources and successfully demonstrate segmentation and classification capabilities without GPU accelerators.

1 Introduction

There are many useful applications of image segmentation in retail. Segmentation can be used to categorize different products within an image, allowing retailers to manage their inventory, keep track of stock levels, and identify which products are in demand. Customers can also take pictures of products they like from their smartphones, and AI algorithms can segment the image to identify the specific item, making it easier for customers to find and purchase similar products. With segmentation, customers can even skip having to manually scan their items at the checkout as their purchases can be gathered from image segmentation. That said, most segmentation models are computationally costly and are unable to be deployed onto hardware devices with limited computation. For this reason, it is important to understand how segmentation models can be modified so that they may be deployed to hardware devices.

EfficientViT helps us to address this issue. The family of lightweight vision transformer models introduced by Xinyu et al. reduces the memory inefficient operations commonly found in most vision transformer models (1). This allows the model to perform well on devices with limited computing in comparison to more demanding vision transformers. We use the EfficientViT for Segment Anything Model to segment retail items, later classifying them for potential applications (2). We then deploy this functionality onto a personal desktop device with limited computing for testing.

2 Relevant Works

Traditionally, for any image segmentation problem, there are two main approaches. First, is interactive segmentation, where any class of objects may be segmented but relies on human supervision and guidance for iterative mask refinement. Secondly, is automatic segmentation, where a set of predefined classes of objects are segmented but require a significant amount of annotated objects. Meta AI introduced the Segment Anything Model (SAM) for the generalization of the two approaches (2).

SAM took the idea of "prompting" from the field of natural language process and applied it to image segmentation. A prompt is any information that specifies what to segment in an image (e.g.

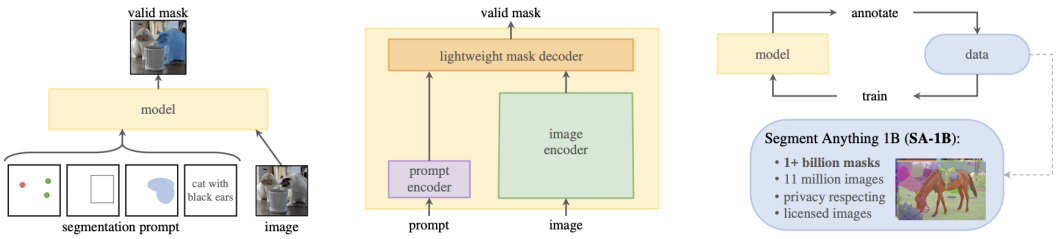

Figure 1: Overview workflow of SAM from Figure 1 of the Segment Anything paper

foreground/background points, bounding boxes, outlines, masks, text, etc.). The output is a number of valid masks along with confidence scores.

Along with the image segmentation model was the Segment Anything 1-Billion mask dataset (SA-1B), which the model trained on and is publicly available for research purposes. It contains over 11 million quality images, which was state-of-the-art at its time of release.

EfficientViT, is a new family of vision transformer models proposed by the MIT HanLab that are efficient for high-resolution dense prediction vision tasks (1). Its novelty comes from replacing the softmax attention module with a lightweight multi-scale linear attention module with hardware efficient operations. It was shown that applying EfficientViT with SAM allowed a significant speedup while maintaining quality performance.

Figure 2: Visualization from Figure 7 of EfficientViT paper demonstrating EfficientViT vs. ViT-H used in SAM

3 Methods

There are three main steps for our EfficientViT for Retail model: 1. segmentation, 2. partition, and 3. classification. In our study, our model has been trained to classify fruits and vegetables. Other classification labels and retail products could be added in the future as our next step. Our workflow diagram can be found in Figure 3.

First, when given an image of retail (fruits and vegetables), our model runs the EfficientViT model proposed by Xinyu et al. to run the SAM model efficiently. This means, given an image that contains different retail products and/or non-retail objects, the EfficientViT will segment every part of the image regardless of whether they are retail products or not. Given an W x H sized image, this first segmentation layer returns $S_1, S_2, ..., S_N$ of (W, H) segment tensors, where N is the number of segments identified by EfficientViT. For every (i, j) in S_x, (i, j) is True if the pixel (i, j) in the original image is part of the segment x, and False otherwise. We use these segment tensors to run our next step, parsing.

Next, from N segment tensors the model generated in the first step, we can generate $P_1, P_2, ..., P_N$ parsed images where each image only includes the portion of the original image that belongs to that specific segment. In order to achieve this goal, in our parsing step, we iterate through N segment tensors to generate N parsed images for all the segments. Specifically, given an original image O with the dimension of W x H, segment tensor S_x would lead to a parsed image P_x where each pixel

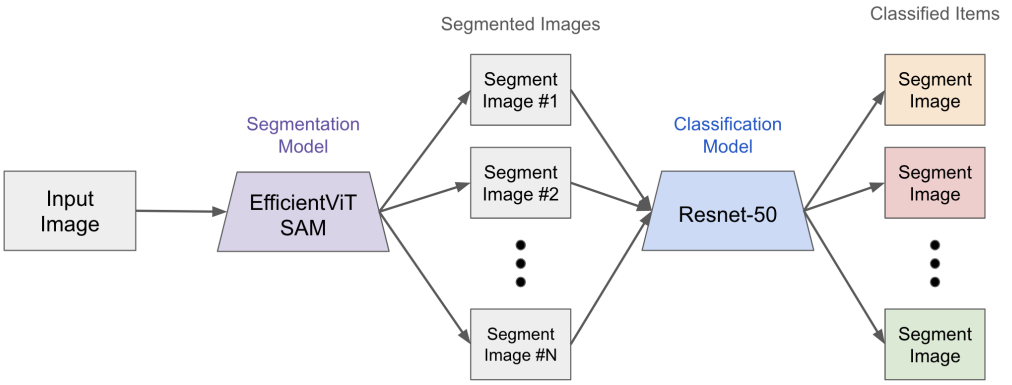

Figure 3: Diagram showing how images are segmented and classified

$P_x[i][j]$ in the parsed image is defined by the following formula:

$$P_x[i][j] = \begin{cases} O[i][j] & \text{if } S_x[i][j] = \text{True} \\ \text{opaqueness } 0 & \text{if } S_x[i][j] = \text{False} \end{cases} \tag{1}$$

Then, each parsed image P_x is saved as "[original image name]_segment_[x].png" in the segmented image folder. These images are then processed to go through the final layer of our model, classification.

As the final step, the parsed images are classified into their labels with an image classification. To do so, we first needed to train our image classification model. In our model, we inherited the architecture of the ResNet-50 model as the backbone of our classification layer (3). ResNet-50, a widely used visual/image classification task in the Computer Vision ML field, consists of multiple layers including Zero Padding, Convolutional, Batch Norm, ReLu, etc. The model has proven to perform superbly on different domains of image classification, such as achieving 80.4% top-1 accuracy at resolution 224x224 on ImageNet-1K (4). To train this backbone image classification model into our specific task—retail products classification—we used the "Fruits and Vegetables" dataset (5). This dataset consisted of 36 classes of fruits and vegetables in total: "banana", "apple", "pear", "cucumber", "carrot", etc.. The dataset included the train/test/validation split, where the training set consisted of 100 images, the test set with 10 images, and the validation set with 10 images per class. Because of the limitation of our dataset, our model currently is limited to classifying the segmented images only into one of these 36 classes.

With this trained classification layer, we can label every partitioned image into one of the 36 classes as described above. However, among the segmented images may contain non-retail objects, incorrectly-segmented retail objects, and retail objects that do not belong to one of the 36 labels (such as durian). To resolve this issue, we introduced the confidence cutoff, where the partitioned images get through the classification later only when the model labels the image with the confidence higher than the confidence cutoff. Since our classification layer employs a multi-class classification using ResNet-50, we can extract the level of confidence our classification model by using SoftMax on the last hidden state of 36 nodes. Our study used the confidence cutoff of 0.78 to produce the results, but the confidence cutoff also serves as a hyperparameter that could be tested and examined.

4 Results

Following successful deployment of our EfficientViT SAM model, we are able to segment images containing retail items. That being said, we find that our EfficientViT may sometimes decompose a single retail item into numerous pieces, making it difficult for our classifer to properly classify items. Figure 4 shows an example of this. While EfficientViT is able to segment the three fruits, it also happens to segment the stem of the banana into an image that the classifier classifies as "ginger". That being said, our SAM model is able to capture the three fruits that present.

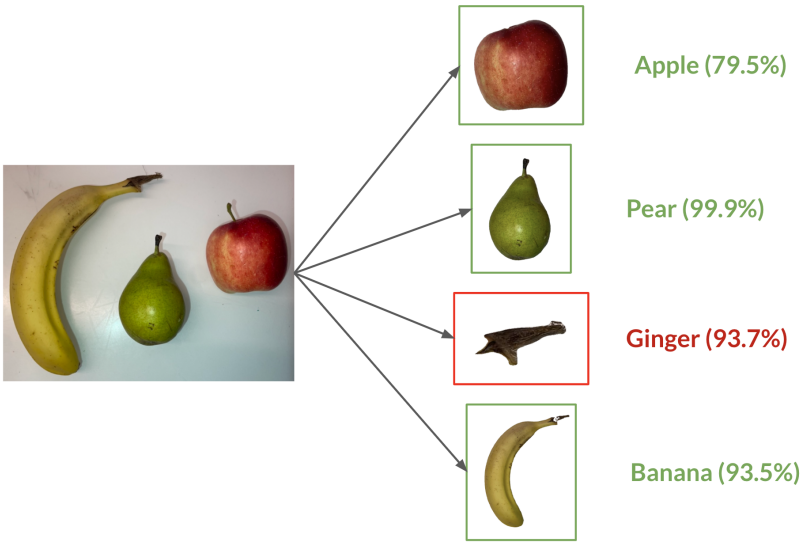

Figure 4: An example of an image containing multiple fruits being segmented and classified. The numbers in the parentheses indicate the confidence with which our classification model predicted each segmented image

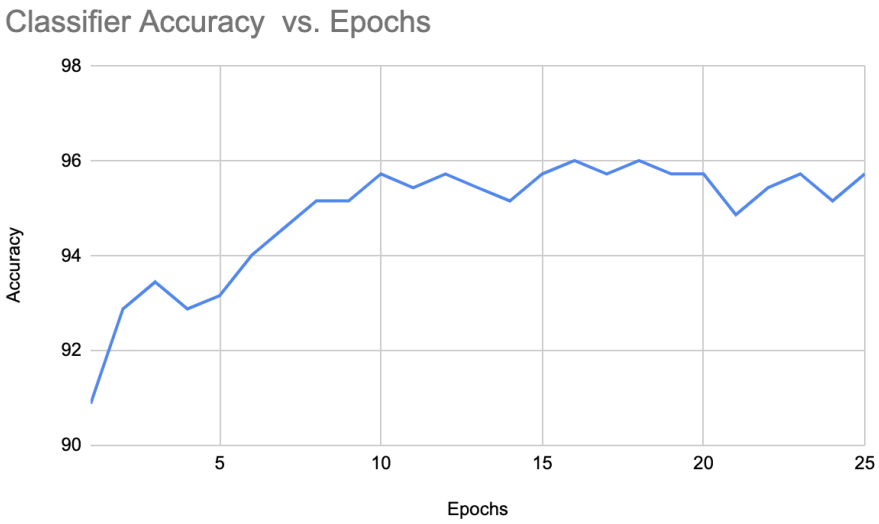

Figure 5: A graph of the Resnet-50 accuracy scores across epochs trained on a dataset of fruits and vegetables

In regards to our Resnet-50 classifer, we find that after 25 epochs of training as shown in Figure 5, the top-1 accuracy of the model is 95.73% when evaluated against the testing split of our fruits and vegetables dataset. Despite this accuracy, we find that it can have trouble classifying fruits positioned in odd orientations or have been partially hidden by other objects. This is likely a result of a lack of such cases in the training data. Aside from these edge cases, our Resnet classifier tends to classify fruits and vegetables with high degress of confidence, as can also be seen in Figure 4.

5 Conclusion and Future Works

In this paper, we found that combining EfficientViT model and image classification model such as ResNet-50 could be used to produce promising visual segmentation and labeling results on real-world application, such as retail (1). More specifically, with our classification layer that achieved the

validation accuracy of 95.73% on labeling images of 36 different fruits and vegetables, our model can conduct segmentation, parsing, and classification tasks at one go with no human interference.

Due to time and computing constraints, we understand there are many potential improvements that could be made to our model. For future works, we hope to train our dataset that contains much more diverse classes of retail products besides the 36 labels used in our study. These may include other products people can commonly find in convenience/grocery stores such as drinks, snacks, and medicine. Moreover, this image segmentation and classification model could be applied to other domains where efficient, speedy inference could be of great help, such as medical field (6). Finally, we can speed up the inference and make the model more efficient by utilizing different efficiency techniques, such as pruning and quantization, on the ResNet-50 classification layers (7).

References

[1] H. Cai, C. Gan, and S. Han, "Efficientvit: Enhanced linear attention for high-resolution low-computation visual recognition," *arXiv preprint arXiv:2205.14756*, 2022.

[2] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[4] R. Wightman, H. Touvron, and H. Jégou, "Resnet strikes back: An improved training procedure in timm," *arXiv preprint arXiv:2110.00476*, 2021.

[5] K. Seth, "Fruits and vegetables image recognition dataset." `https://www.kaggle.com/datasets/kritikseth/fruit-and-vegetable-image-recognition/`, 2020.

[6] Y. Xu, J.-Y. Zhu, I. Eric, C. Chang, M. Lai, and Z. Tu, "Weakly supervised histopathology cancer image segmentation and classification," *Medical image analysis*, vol. 18, no. 3, pp. 591–604, 2014.

[7] T. Liang, J. Glossner, L. Wang, S. Shi, and X. Zhang, "Pruning and quantization for deep neural network acceleration: A survey," *Neurocomputing*, vol. 461, pp. 370–403, 2021.