Background Removal

A. Malonosova

May 8, 2022

1 Related Works

To remove background we need to find foreground objects (or objects) and remove all others. The problem of finding such objects is famous as Salient Object Detection. It aims at localizing the most visually attractive objects in an image. Nevertheless, significant number of models for this task is adapted from the other task – Semantic Segmentation. The task of Semantic Segmentation is to split the image to several "objects", i.e. assign each pixel its own class. Note, that if image contains only one foreground object and background is one object, then the problem of Semantic Segmentation is almost equivalent to Salient Object Detection. The example of such picture is a selfie with simple background (e.g. sky).

The general approach for semantic segmentation problem is neural networks based on encoder and decoder. The encoder is usually a network that map the initial image to some latent space. Sometimes, the layers for encoder are taken from pre-training deep neural networks for classification, like ResNet or VGG. The task of the decoder is to project the latent vector obtained by the encoder onto the image space with classified pixels. The example of such neural network is Fully Convolution Network (FCN) (see [1]). The authors of this problem take pre-trained ALexNet convolution layers and uses them as Encoder. For decoder, they take simple convolution layers. Note, that this architecture was adopted for SOD problem on video in work [2]. They trained convolution and deconvolution layers. They train this layers together on Cross-Entropy Loss. Another generalization of FCN represented in the work [3]. They based their architecture on VGG blocks and improved this through so-called guidance blocks to embed edge prior knowledge into hierarchical feature maps for effective feature representations. They minimize the weighted sum of cross-entropy loss and IOU boundary loss. Note, that pretrained AlexNet, VGG and ResNet models are available in PyTorch. Also, an example of using of FCN model is presented in `https://gist.github.com/khanhnamle1994/e2ff59ddca93c0205ac4e566d40b5e88`.

Also, let us note specific work [4]. Their authors proposed so-called CascadePSP model. This model does not implement segmentation, but tries to improve its quality. Their model takes initial image and corresponding mask, constructed by other model, passes them to Encoder-Decoder Architecture with skip-connections and returns improved mask. Encoder represented by PSPNet and ResNet pre-trained layers, when decoded represented by simple deconvolution layers. This model demonstrates significant quality improvement for PSPNet, FCN, DeepLab and RefineNet models in IoU metric. The code for CascadePSP is available `https://github.com/hkchengrex/CascadePSP`. It is based on PyTorch.

Another class of neural network for semantic segmentation and salient object detection is region-based methods. So, in work [5], R-CNN implements selective search to find the

significant number of objects and after that computes their CNN features. Finally, it classifies each region using class-specific SVMs. This neural network can be based on different famous deep neural network for classification like AlexNet, ResNet or VGG. Note, that this approach was generalized in work [6] for SOD problem. The architecture of their deep feature based model for visual saliency consists of one output layer and two fully connected hidden layers on top of three deep convolutional neural networks. They extract multiscale features for each image region with a deep convolutional neural network originally trained over the ImageNet dataset using Caffe, an open source framework for CNN training and testing.

Further, let us consider specific works for SOD problem. One of such works is [7]. The authors focus on the quality of bounders of the detected object, but not on internal space of this object. Their neural network BASNet represents stacked two Encoder-Decoder networks with residual blocks. The first module is so-called Predict Model based on architectures Unet and Segnet. Its purpose to predict result. The second model is Refine Module that improves quality of prediction. They train their model on sum of three terms: binary cross-entropy, SSIM and smooth version of IOU loss. SSIM is originally proposed for image quality assessment. It captures the structural information in an image. The code is available `https://github.com/xuebinqin/BASNet`. Model is written on PyTorch.

At the same time, work [10] represents effictive using of deep features generated by proposed IndexNet model. The proposed IndexNet dynamically predicts indices for individual local regions, conditional on the input local feature map itself. The predicted indices are further utilized to guide the downsampling in the encoding stage and the upsampling in corresponding decoding stage. The code for model is available `https://github.com/open-mmlab/mmediting/tree/master/configs/mattors/indexnet`.

Another work for salient object detection is the recent work [8]. They also based their model on Unet architecture and named it as U^2-Net. It represents is a two-level nested U-structure. Note, that the way of stacking different architectures and, in particular, UNet is well known practice. In fact, they uses a lot of UNet models and construct from them new UNet where each block is UNet-like. The authors state that their architecture has the following advantages. Firstly, one can apply high-capacity convolutional neural networks (CNNs) to bottom-up region proposals in order to localize and segment objects. Secondly, when labeled training data is scarce, supervised pre-training for an auxiliary task, followed by domain-specific fine-tuning, yields a significant performance boost. It is the recent work and they states inspired results in comparison with earlier networks. The code for this paper is presented in `https://github.com/NathanUA/U-2-Net`. Also, there is code of using of this model for BackGround removal `https://github.com/danielgatis/rembg`. Also, note here another using of this model: `https://github.com/dennisbappert/u-2-net-portrait` (U^2-Net portrait matting) and `https://github.com/xuebinqin/U-2-Net` (U^2-Net portrait matting, Background Removal and etc.).

Also, note several approaches for related problem Background Removal - Portlet Matting. The first considered model is MODNet [11]. The code for this model is available `https://github.com/ZHKKKe/MODNet`. MODNet predicts portrait semantics and boundary details, and after that, units this. Authors states that their network is fast enough for online using. Also, note, that there is the following project for Real-Time High-Resolution Background Matting `https://github.com/PeterL1n/BackgroundMattingV2`. It is an official repository for [12]. Their model based on DeepLabV3 neural network.

Note, that described above approaches trained in a supervised way. Further, let us consider weakly-supervised approach. The particular case of weakly-supervised salient ob-

ject detection is WSS model, proposed in work [9]. Their model contains two parts: FCN and Foreground Inference Net (FIN), new network for a generic salient object detection. The WSS trains in two stages. On the first stage FCN and FIN are trained together for image categorization (supervised part). After that, FIN is trained for saliency prediction (unsuperives part) for specific loss. Though supervised method demonstrate better performance, the difference is not significant.

2 Datasets

1. *PASCAL-S*: This dataset has 850 natural images which are generated from the PASCAL VOC segmentation challenge. The ground truths labeled by 12 experts contains both pixel-wise saliency and eye fixation

2. *DUT-OMRON*: This is a large dataset with 5168 high quality images. Each image in this dataset has one or more salient objects and a cluttered background. Therefore, this dataset is more difficult and challenging, which provides more space for improvement for the research of saliency detection.

3. *ECSSD*: This dataset contains 1000 natural and complex images with pixel-wise ground truth annotations and these images are manually selected from the Internet.

4. *SOD*: This dataset has 300 images, and it was originally designed for image segmentation. It is challenging be- cause many images have multiple objects which with low contrast or touching the image boundary.

5. *DUTS* dataset consists of two parts: DUTS-TR and DUTS-TE. As mentioned above we use DUTS-TR for training. Hence, DUTS-TE, which contains 5019 images, is selected as one of our evaluation dataset.

6. *HKU-IS* contains 4447 images with multiple foreground objects.

7. *ECSSD* contains 1000 structurally complex images and many of them contain large foreground objects.

8. *SED* comprises a single-object subset and a two- object subset; each has 100 images with mask annotations.

9. *MSRA10K*, also known as THUS10K, contains 10,000 images selected from MSRA-A and covers all the images in ASD. Due to its large scale, MSRA10K is widely used to train deep SOD models.

10. *SOS* is created for SOD subitizing. It contains 6,900 images (training set: 5,520, test set: 1,380). Each image is labeled as containing 0, 1, 2, 3 or 4+ salient objects. *MSO* is a subset of SOS-test, covering 1,224 images. It has a more balanced distribution of the number of salient objects. Each object has a bounding-box annotation.

11. *ILSO* contains 1,000 images with precise instance- level annotations and coarse contour labeling.

12. *XPIE* has 10,000 images with pixel-wise labels. It has three subsets: Set-P has 625 images of places-of-interest with geographic information; Set-I 8,799 images with object tags; and Set-E 576 images with eye-fixation records.

13. *SOC* consists of 6,000 images with 80 common categories. Half of the images contain salient objects, while the remaining have none. Each image containing salient objects is annotated with an instance-level ground-truth mask, object category, and challenging factors. The non- salient object subset has 783 texture images and 2,217 real- scene images.

14. *COCO-CapSal* is built from COCO and SALICON. Salient objects were first roughly localized using the mouse-click data in SALICON, then precisely annotated according to the instance masks in COCO. The dataset has 5,265 and 1,459 images for training and testing, respectively.

15. *HRSOD* is the first high-resolution dataset for SOD. It contains 1,610 training and 400 testing images collected from websites. Pixel-wise ground-truths are provided.

3 Our Approach

We have chosen for our Background Removal project two models: U^2-Net [8] and CascadePSP [4].

U^2-Net presents specific UNet nested architecture. So, this model be constructed from UNet-like blocks - RSU (**R**e**S**idual textbfU-block). Each such block contain a input convolution layer, encoder-decoder architecture from UNet and a residual connection. So, the design of U2-Net allows having deep architecture with rich multi-scale features. Model U^2-Net in original paper was trained on *DUT-TR* dataset that contains 10553 images. The training process according to article requires about 120 hours. So, because of limited time we taken pre-trained model from the repository `https://github.com/xuebinqin/U-2-Net`.

Before running U^2-Net model, all images are resized to shape (128, 128) and normalized.

After U^2 net work we take the obtained mask with initial image and pass its through CascadePSP model. This model significantly improves quality of obtained result. The CascadePSP model was taken pre-trained too. We used the module from the repository `https://github.com/hkchengrex/CascadePSP`.

We tested quality on datasets *DUTS*, *DUT-OMRON* and *ECSSD*. We researched quality on three metrics: IOU, Presicion, Recall and F_β for $\beta = 0.3$.

References

[1] **Jonathan Long, Evan Shelhamer, Trevor Darrell** Fully Convolutional Networks for Semantic Segmentation // arXiv preprint (2015). Url: https://arxiv.org/pdf/1411.4038.pdf.

[2] **Wenguan Wang, Jianbing Shen, Ling Shao** Video Salient Object Detection via Fully Convolutional Networks // arXiv preprint (2017). Url: https://arxiv.org/pdf/1702.00871.pdf.

[3] **Zhengzheng Tu, Yan Ma, Chenglong Li, Jin Tang, Bin Luo** Edge-guided Non-local Fully Convolutional Network for Salient Object Detection // arXiv preprint (2019). Url: https://arxiv.org/pdf/1908.02460.pdf.

[4] **Ho Kei Cheng, Jihoon Chung, Yu-Wing Tai, Chi-Keung Tang** CascadePSP: Toward Class-Agnostic and Very High-Resolution Segmentation via Global and Local Refinement // arXiv preprint (2020). Url: https://arxiv.org/abs/2005.02551.

[5] **Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik** Rich feature hierarchies for accurate object detection and semantic segmentation // arXiv preprint (2014). Url: https://arxiv.org/pdf/1311.2524.pdf.

[6] **Guanbin Li, Yizhou Yu** Visual Saliency Based on Multiscale Deep Features // arXiv preprint (2015). Url: https://arxiv.org/pdf/1503.08663.pdf.

[7] **Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, Martin Jagersand** BASNet: Boundary-Aware Salient Object Detection // CVPR 2019. Url: https://openaccess.thecvf.com/content_CVPR_2019/papers/Qin_BASNet_Boundary-Aware_Salient_Object_Detection_CVPR_2019_paper.pdf.

[8] **Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R. Zaiane, Martin Jagersand** U^2-Net: Going Deeper with Nested U-Structure for Salient Object Detection // arXiv preprint (2022). Url: https://arxiv.org/pdf/2005.09007.pdf.

[9] **Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, Xiang Ruan** Learning to Detect Salient Objects with Image-level Supervision // CVPR 2017. Url: https://openaccess.thecvf.com/content_cvpr_2017/papers/Wang_Learning_to_Detect_CVPR_2017_paper.pdf.

[10] **Hao Lu, Yutong Dai, Chunhua Shen†, Songcen Xu** Indices Matter: Learning to Index for Deep Image Matting // arXiv preprint (2019). Url: https://arxiv.org/pdf/1908.00672.pdf.

[11] **Zhanghan Ke, Jiayu Sun, Kaican Li, Qiong Yan, Rynson W.H. Lau** MODNet: Real-Time Trimap-Free Portrait Matting via Objective Decomposition // arXiv preprint (2020). Url: https://arxiv.org/pdf/2011.11961.pdf.

[12] **hanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian Curless, Steve Seitz, Ira Kemelmacher-Shlizerman** Real-Time High-Resolution Background Matting // arXiv preprint (2020). Url: https://arxiv.org/pdf/2012.07810.pdf.