Note: Learning What Not to Segment: A New Perspective on Few-Shot Segmentation

sgc

May 1, 2022

1 Introduction

Few-shot learning (FSL) is proposed to address the problem of notorious labeling by building a network that can be generalized to unseen domains with scarce annotated samples available.
And the few-shot segmentation(FSS) is part of the task. Most of works on it are achieved by meta-learning which inevitably introduces bias towards seen classes.
This paper proposed BAM(Base and Meta) which introduces an extra branch learner, and the coarse output of 2 learners are integrated into a accurate prediction.

2 Method

2.1 Base learner

First use a encoder and a convolution block to get the query image's information. And then a decoder is used to enlarge the feature map to yield the prediction.

$$\mathbf{p}_{\mathrm{b}}=\operatorname{softmax}\left(\mathcal{D}_{\mathrm{b}}\left(\mathbf{f}_{\mathrm{b}}^{\mathrm{q}}\right)\right) \in \mathbb{R}^{(1+N_{\mathrm{b}}) \times H \times W},$$

where N_b represents the number of base classes.
The cross-entropy loss of the base learner:

$$\mathcal{L}_{\text{base}}=\frac{1}{n_{\mathrm{bs}}} \sum_{i=1}^{n_{\mathrm{bs}}} \mathrm{CE}\left(\mathbf{p}_{\mathrm{b};i}, \mathbf{m}_{\mathrm{b};i}^{\mathrm{q}}\right),$$

where the m_b^q is the GT, and n_{bs} represents the number of samples in each batch.

2.2 Meta Learner

Given a support set $S = x_s, m_s$ and a query imagex_q, the goal of the meta learner is to segment the objects in x_q that share the same category as the annotation mask m_s under the guidance of S.
The class-related cues v_s:

$$\mathbf{v}_{\mathrm{s}} = \mathcal{F}_{\mathrm{pool}}\left(\mathbf{f}_{\mathrm{m}}^{\mathrm{s}} \odot \mathcal{I}\left(\mathbf{m}^{\mathrm{s}}\right)\right) \in \mathbb{R}^{c},$$

where I reshape m_s into same size of f.
The final prediction results are generated as:

$$\mathbf{p}_{\mathrm{m}} = \operatorname{softmax}\left(\mathcal{D}_{\mathrm{m}}\left(\mathcal{F}_{\mathrm{guidance}}\left(\mathbf{v}_{\mathrm{s}}, \mathbf{f}_{\mathrm{m}}^{\mathrm{q}}\right)\right)\right) \in \mathbb{R}^{2 \times H \times W},$$

where D_m is the decoder network, and $F_{guidence}$ is an essential module of FSS that passes the annotation information from the support branch to the query branch to provide specific segmentation cues.
The BCE loss:

$$\mathcal{L}_{\mathrm{meta}} = \frac{1}{n_{\mathrm{e}}} \sum_{i=1}^{n_{\mathrm{e}}} \mathrm{BCE}\left(\mathbf{p}_{\mathrm{m};i}, \mathbf{m}_{i}^{\mathrm{q}}\right),$$

where n_e represents the training episodes in each batch.

3 Ensemble

First integrate the foreground probability maps generated by the base learner to obtain the prediction of the background region relative to the few-shot task through just sum up p_b^i.
Compute Φ the adjustment factor by:

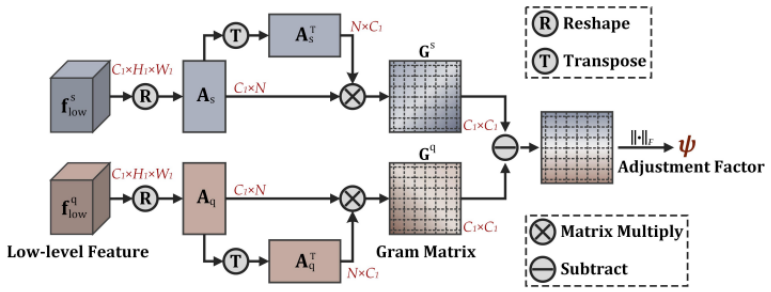

Figure 3. The calculation process of the adjustment factor ψ for the low-level features $\mathbf{f}_{\mathrm{low}}^{\mathrm{s}}$ and $\mathbf{f}_{\mathrm{low}}^{\mathrm{q}}$.

The final prediction:

$$\mathbf{p}_{\mathrm{f}}^{0}=\mathcal{F}_{\text{ensemble}}\left(\mathcal{F}_{\psi}\left(\mathbf{p}_{\mathrm{m}}^{0}\right),\mathbf{p}_{\mathrm{b}}^{\mathrm{f}}\right),$$

$$\mathbf{p}_{\mathrm{f}}=\mathbf{p}_{\mathrm{f}}^{0}\oplus\mathcal{F}_{\psi}\left(\mathbf{p}_{\mathrm{m}}^{1}\right),$$

The total loss:

$$\mathcal{L}=\mathcal{L}_{\text{final}}+\lambda\mathcal{L}_{\text{meta}},$$

$$\mathcal{L}_{\text{final}}=\frac{1}{n_{\mathrm{e}}}\sum_{i=1}^{n_{\mathrm{e}}}\mathrm{BCE}\left(\mathbf{p}_{i}^{\mathrm{q}},\mathbf{m}_{i}^{\mathrm{q}}\right),$$

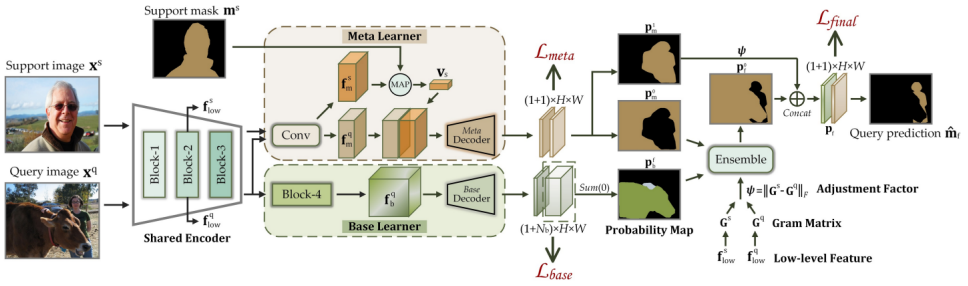

Figure 2. Overall architecture of the proposed BAM, which is composed of three essential components: a base learner, a meta learner, and an ensemble module. In each training episode, the two learners extract the features of input image pairs $(\mathbf{x}^s, \mathbf{x}^q)$ with a shared encoder and make predictions for the specific base category c (note that c denotes the novel category in the meta-testing phase) and the remaining base categories, respectively. Then, the coarse predictions are fed to the ensemble module along with an adjustment factor ψ to suppress the falsely activated regions of base categories, further producing accurate segmentation results. For ease of understanding, we present the probability maps in the form of segmentation masks, but they are actually two-dimensional floating-point matrices, *i.e.*, $\mathbf{p} \in [0,1]^{H\times W}$. MAP represents the masked average pooling operation [67].

3.1 K-Shot Setting

Φ represents the affinity between the support samples and query samples, thus Φ can be used to guide to ensemble.

$$\eta=\text{soft}\max\left(\mathbf{w}_{2}^{\mathsf{T}}\mathrm{ReLU}\left(\mathbf{w}_{1}^{\mathsf{T}}\psi_{\mathrm{t}}\right)\right)\in\mathbb{R}^{K},$$

the η is the weight of support samples in K-Shot.

3.2 4.5. Extension to Generalized FSS

Use a threshold τ to generate holistic segmentation prediction:
ıı be formulated as:

$$\hat{\mathbf{m}}_{\mathrm{g}}^{(x,y)}=\begin{cases} 1 & \mathbf{p}_{\mathrm{f}}^{1;(x,y)} > \tau \\ \hat{\mathbf{m}}_{\mathrm{b}}^{(x,y)} & \mathbf{p}_{\mathrm{f}}^{1;(x,y)} \leq \tau \text{ and } \hat{\mathbf{m}}_{\mathrm{b}}^{(x,y)} \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\hat{\mathbf{m}}_{\mathrm{b}}=\arg\max\left(\mathbf{p}_{\mathrm{b}}\right) \in \{0,1,...,N_{\mathrm{b}}\}^{H\times W},$$

4 Results

In quantity, BAM outperforms the advanced FSS models with a considerable margin and sets new state-of-the-arts under all settings.
In quality, the falsely activated targets of base classes are significantly suppressed.

5 Ablation Study

1, Train the two learners separately works better than jointly.
2, B2 yields better trade-off result in generating Φ.
3, Bounding box annotation works well too.

6 Summary

BAM successfully alleviated the bias of existing meta-learning method in FSS models. The core idea is leveraging the base learner to identify the confusable (base) regions in the query images and further refine the prediction of the meta learner. This method set the new benchmarks.

7 Extra knowledge

1, Meta-learning: Learn to cope with new data after trained in many other class information."https://www.zhihu.com/question/264595128"

References