CALSEG: IMPROVING CALIBRATION OF MEDICAL IMAGE SEGMENTATION VIA VARIATIONAL LABEL SMOOTHING

ICASSP 2024 KOREA

Xutao Guo[1,2], Yanwu Yang[1,2], Chenfei Ye[3,2], Guoqing Cai[1], Ting Ma[1,2,3,*]

1. School of Electronics and Information Engineering, Harbin Institute of Technology at Shenzhen, Shenzhen, China
2. Peng Cheng Laboratory, Shenzhen, China
3. International Research Institute for Artifcial Intelligence, Harbin Institute of Technology at Shenzhen, Shenzhen, China

Abstract

In practical medical image segmentation tasks, ensuring confidence calibration is crucial. However, medical image segmentation typically relies on hard labels (one-hot vectors), and when minimizing the cross-entropy loss, the model's softmax predictions are compelled to align with hard labels, resulting in over-confident predictions. To alleviate above problems, this study proposes a novel framework on calibration of medical image segmentation, called CALSeg.

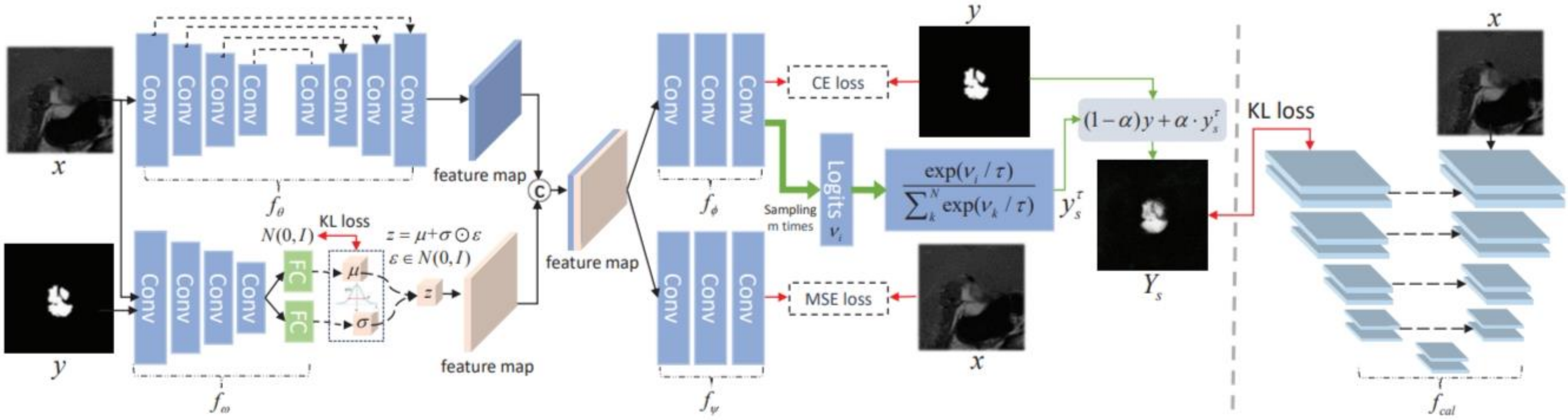

Fig. 1: Overall of CALSeg framework. Left: the Variational Label Smoothing (VLS) estimates the soft labels Y_s. Right: we combine the estimated soft labels with the hard labels for calibration model training.

Method

Variational Label Smoothing

Based on Bayesian theory, we can use hard labels to estimate the corresponding potential soft labels. First, learn a probability model to capture the underlying joint distribution p(z|x,y) between the images x and corresponding hard labels y. The hard label y can be considered as generated from the conditional distribution p(y|z). Therefore, we can sample multiple times from p(y|z) to generate soft labels ys. Due to the difficulty of solving the integral computation, variational inference (VI) is used to compute the posterior distribution p(z|x, y). VI introduces a fixed-form distribution q(z|x, y, w) parameterized by w to approximate the true posterior distribution p(z|x, y).

Training:

The objective of VI is to minimize the reverse KL divergence between distributions p(z|x, y) and q(z|x, y). This can be expressed as follows:

$$KL[q(z|x,y)||p(z|x,y)] = \mathbb{E}_{q(z|x,y)}[\log \frac{q(z|x,y)}{p(z|x,y)}]$$

$$\mathcal{L}(x,y;\phi,\psi,\omega) = \mathbb{E}_{q_\omega(z|x,y)}[\log p_\phi(y|z) \\ \log p_\psi(x|z)] - KL[q_\omega(z|x,y)||p(z)].$$

Sampling:

After training, each sample is subjected to VLS sampling m times, and the obtained m sampled probability predictions are averaged to generate the corresponding soft labels ys. However, there may be classification inconsistencies between ys and the original hard labels. To address this, we combine the original hard label

$$Y_s = (1-\alpha)y + \alpha y_s^\tau(x)$$

Experiments

Table 1: The calibration performance (ECE, MCE) and the discriminative performance (DICE) obtained by the different models across two medical image segmentation benchmarks.

Methods	ECE ↓		MCE ↓		DICE ↑			
	WMH	Brats13	WMH	Brats13	WMH	Brats13		
						WT	TC	ET
Baseline[5]	0.141	0.138	0.351	0.302	79.97	73.79	48.44	51.74
TS[6]	0.074	0.082	0.171	0.157	79.95	73.79	48.44	51.74
STS[10]	0.063	0.073	0.136	0.129	79.91	73.79	48.44	51.74
Baye-UNet[12]	0.104	0.109	0.276	0.256	79.71	72.03	50.09	49.27
Prob-UNet[14]	0.122	0.101	0.283	0.213	80.02	76.81	50.89	51.51
DeepEnsemble[15]	0.118	0.112	0.264	0.242	80.04	75.77	51.19	**53.33**
LSR[16]	0.090	0.093	0.178	0.202	80.00	76.85	49.39	52.95
SVLS[17]	0.058	0.037	0.105	0.074	78.12	71.66	46.45	46.01
Self-KD[20]	0.075	0.053	0.131	0.198	80.53	77.57	50.62	52.01
MarginLoss[7]	0.098	0.089	0.211	0.167	79.91	75.15	46.88	50.24
CALSeg	**0.038**	**0.017**	**0.072**	**0.037**	**81.12**	**79.59**	**52.73**	53.01

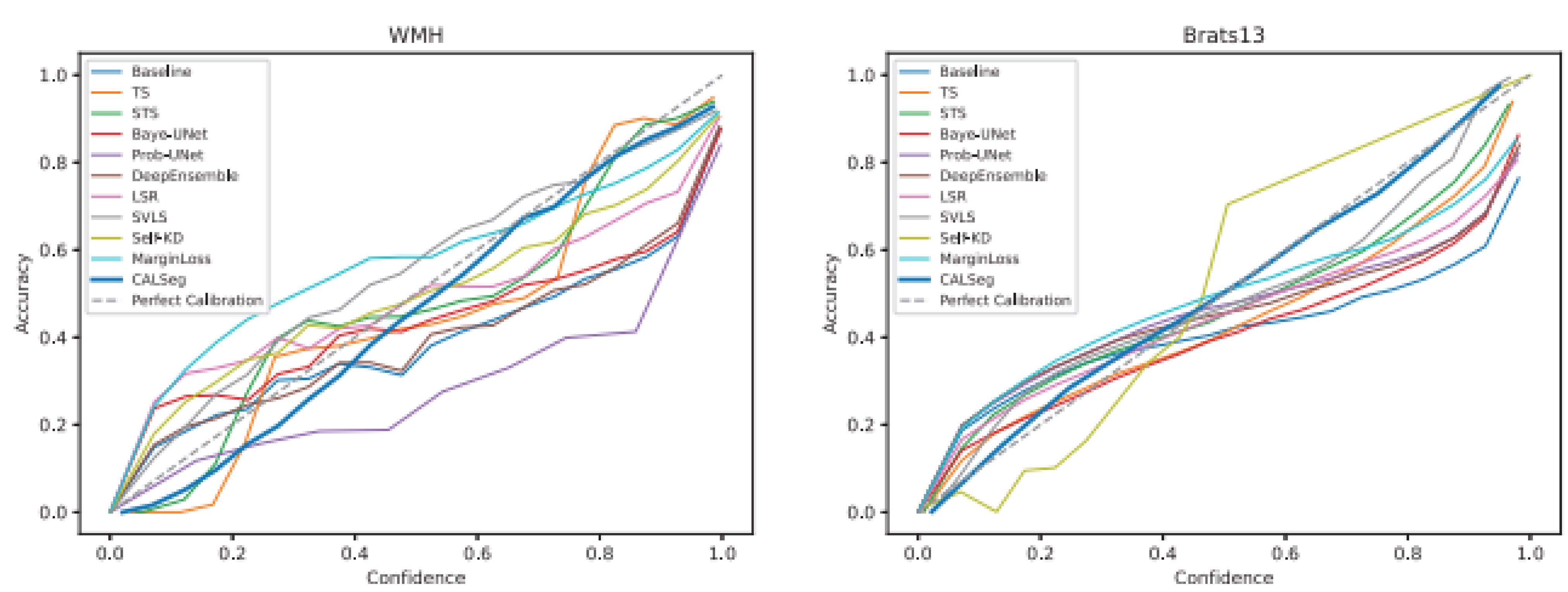

Fig. 2: Reliability diagrams showing calibration between confidence and accuracy for different methods.