Robust T-Loss for Medical Image Segmentation

Alvaro Gonzalez-Jimenez, Simone Lionetti, Philippe Gottfrois, Fabian Groger, Marc Pouly, and Alexander A. Navarini

MICCAI (2023)

Problem

Medical image annotation is:

- Costly
- Time-consuming
- Prone to human bias & labeling errors

Noisy labels affect model training

Previous methods to deal with noisy labels have many hyper-parameters, complex training procedures or modify the network architecture.

Objective

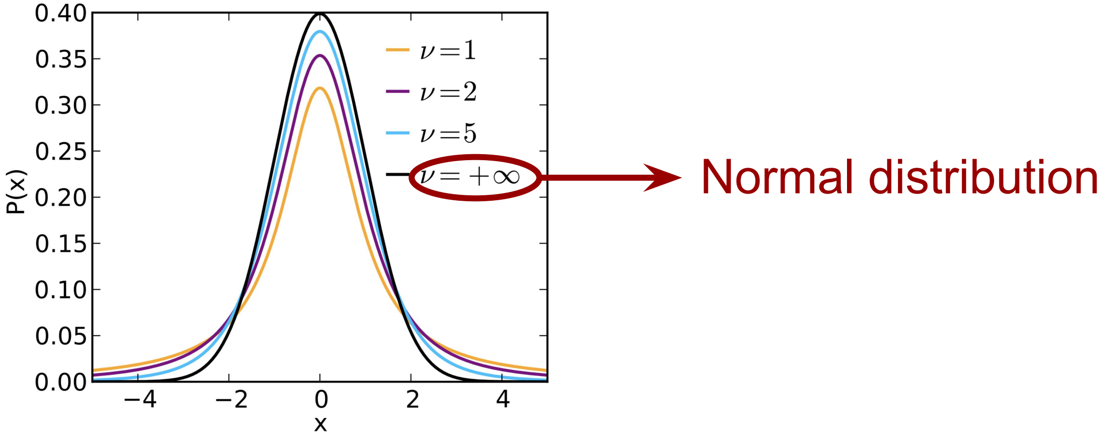
Propose a robust loss function that adaptively learns an optimal tolerance level to label noise → avoid memorizing noisy labels

Method

Hypothesis: Error term follows a Student-t distribution

→ **T-Loss**: based on the negative log-likelihood of the Student-T distribution which:

- Generalizes the normal distribution
- More **robust to outliers and noise** compared to normal distribution



Method

Parameter ν controls **sensitivity to outliers** & is **trained** with the model parameters

$$p(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\Sigma};\nu) = \frac{\Gamma\left(\frac{\nu+D}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{|\boldsymbol{\Sigma}|^{-1/2}}{(\pi\nu)^{D/2}} \left[1 + \frac{(\mathbf{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})}{\nu}\right]^{-\frac{\nu+D}{2}}$$

Assumption: Σ = Identity (pixel annotations independent)

$$\mathcal{L}_{\mathrm{T}} = \frac{1}{N}\sum_{i=1}^{N} -\log p(\mathbf{y}_i|f_{\mathbf{w}}(\mathbf{x}_i),\ \mathbf{I_d};\ \nu)$$

"Robust T-Loss for Medical Image Segmentation", MICCAI (2024)

Method

$$\mathcal{L}_{\mathrm{T}} = \frac{1}{N}\sum_{i=1}^{N} -\log p(\mathbf{y}_i | f_{\mathbf{w}}(\mathbf{x}_i), \mathbf{\Sigma}; \nu)$$

with

$$\begin{aligned} -\log p(\mathbf{y}|\boldsymbol{\mu}, \mathbf{I}_D; \nu) = &-\log \Gamma\left(\frac{\nu + D}{2}\right) + \log \Gamma\left(\frac{\nu}{2}\right) + \frac{D}{2}\log(\pi\nu) \\ &+ \frac{\nu + D}{2}\log\left[1 + \frac{(\mathbf{y} - \boldsymbol{\mu})^2}{\nu}\right] \end{aligned}$$

$$\mathbf{y} = \mathbf{y}_i$$
$$\boldsymbol{\mu} = f_{\mathbf{w}}(\mathbf{x}_i)$$

Error $\delta = |\mathbf{y}_i - f_{\mathbf{w}}(\mathbf{x}_i)|$

1. $\delta \to 0$: loss guided by δ^2(~MSE)
2. Large δ: loss guided by logδ < δ (less penalty than MAE, robust to outliers)

Experiment

- Dataset:
  - ISIC (skin lesion segmentation)
  - Shenzhen (lung segmentation)

  Artificially inject noise to labels (transformations and masks)

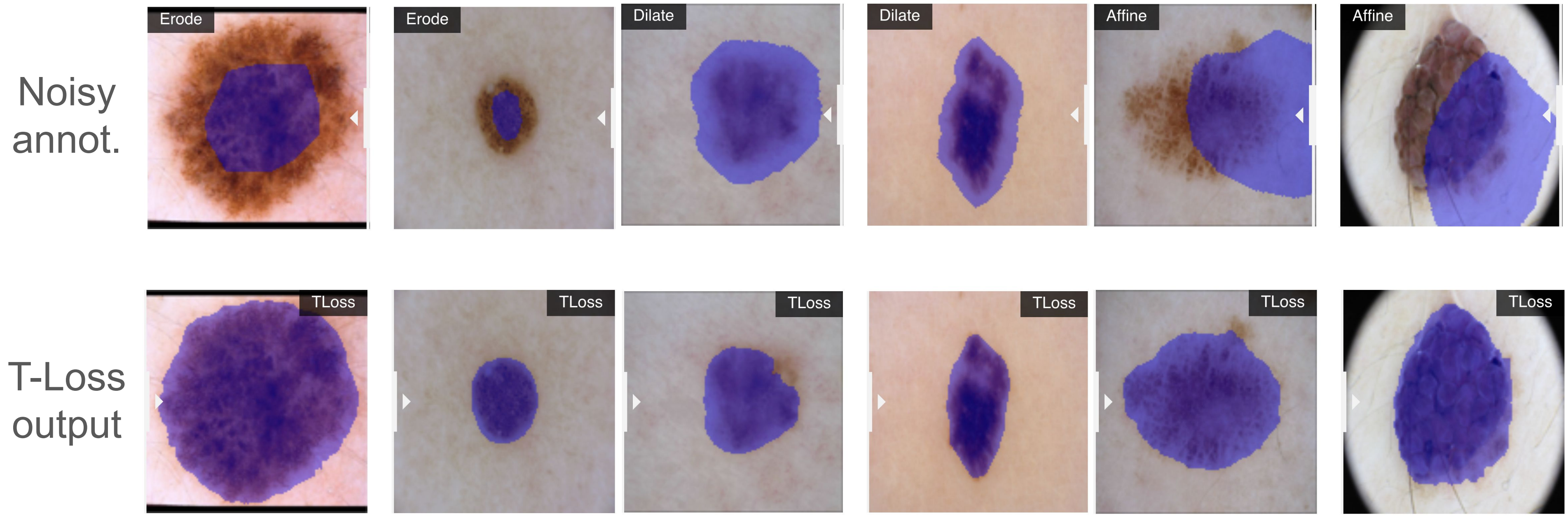
- Model: nn-UNet

Results: Skin lesion segmentation

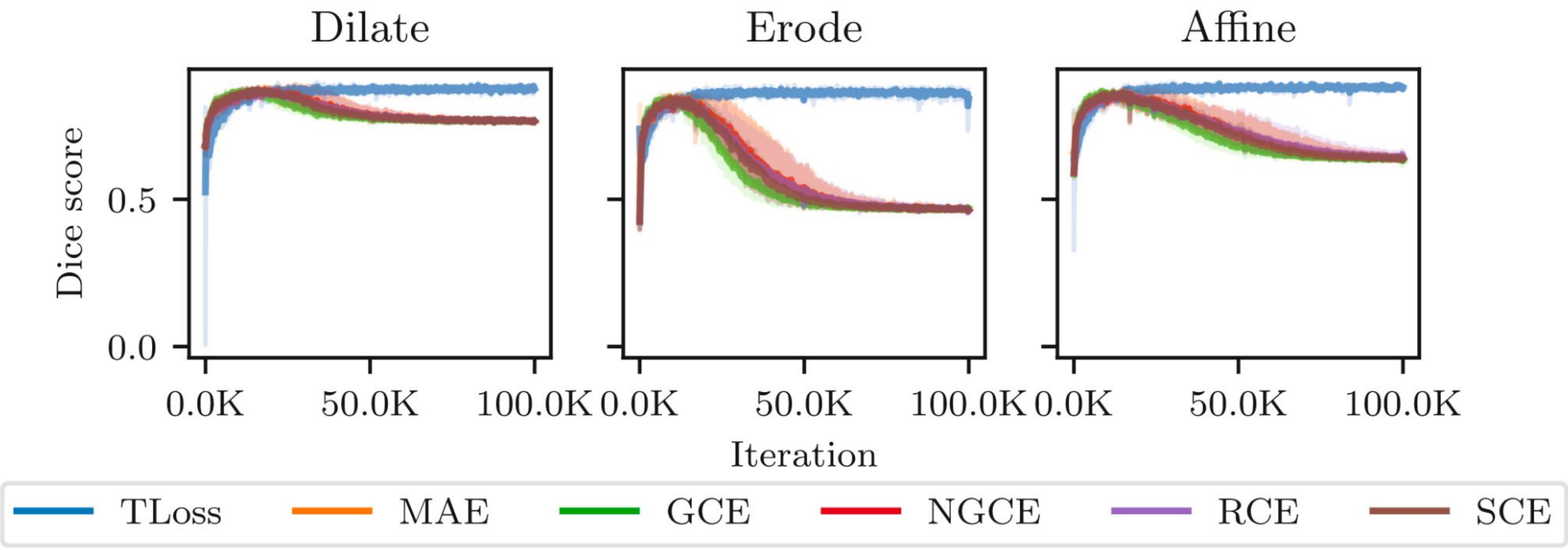
Loss	$\alpha = 0.0$	$\alpha = 0.3$		$\alpha = 0.5$		$\alpha = 0.7$	
		$\beta = 0.5$	$\beta = 0.7$	$\beta = 0.5$	$\beta = 0.7$	$\beta = 0.5$	$\beta = 0.7$
GCE	0.828(7)	0.805(9)	0.785(14)	0.772(15)	0.736(17)	0.743(12)	0.691(22)
MAE	0.826(6)	0.803(7)	0.786(12)	0.771(10)	0.742(14)	0.751(09)	0.698(20)
RCE	0.827(5)	0.802(6)	0.791(11)	0.779(11)	0.745(17)	0.752(10)	0.695(16)
SCE	0.828(6)	0.806(9)	0.793(11)	0.774(12)	0.738(14)	0.756(10)	0.691(20)
NGCE	0.825(6)	0.803(7)	0.788(08)	0.773(11)	0.745(16)	0.745(11)	0.688(19)
NCE+RCE	0.829(6)	0.799(8)	0.792(12)	0.777(11)	0.751(18)	0.746(11)	0.696(13)
NGCE+MAE	0.828(5)	0.802(8)	0.788(13)	0.774(10)	0.741(18)	0.748(11)	0.693(15)
NGCE+RCE	0.827(7)	0.807(7)	0.790(11)	0.776(13)	0.736(17)	0.748(10)	0.689(17)
T-Loss (Ours)	0.825(5)	0.809(6)	$0.804(5)^{\star}$	$0.800(11)^{\star}$	$0.790(5)^{\star}$	$0.788(7)^{\star}$	$0.761(6)^{\star}$

Dice score over different noise noisy label proportion α and noise level □

Results: Visual



Results: Noise memorization



Dice score of training set predictions
→ other methods end up memorizing the noisy labels

Conclusion

T-Loss:

- Can handle outliers in the data
- Has only a single parameter, dynamically optimized

Student-t distribution

$$p(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}; \nu) = \frac{\Gamma\left(\frac{\nu+D}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{|\boldsymbol{\Sigma}|^{-1/2}}{(\pi\nu)^{D/2}} \left[1 + \frac{(\mathbf{y}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})}{\nu}\right]^{-\frac{\nu+D}{2}}$$

Gamma function $\Gamma(n) = (n-1)!$

MSE and log-likelihood

Conditional log-likelihood estimator:
(assuming iid samples)

$$\theta_{ML} = \arg\max_{\theta} \prod_{i=1}^{n} P(y^{(i)}|\mathbf{x}^{(i)};\theta)$$
$$= \arg\max_{\theta} \sum_{i=1}^{n} \log P(y^{(i)}|x^{(i)};\theta)$$

Assuming Gaussian distribution $p(y|\mathbf{x}) = \mathcal{N}(y;\hat{y},\sigma^2)$

Then:

$$\sum_{i=1}^{n} \log p(y^{(i)}|x^{(i)};\theta) = -n\log\sigma - \frac{n}{2}\log(2\pi) - \sum_{i=1}^{n} \frac{\|\hat{y}^{(i)} - y^{(i)}\|^2}{2\sigma^2}$$

$$MSE_{ML} = \frac{1}{m}\left\|\hat{\mathrm{y}} - y\right\|_2^2$$

→ Maximizing log-likelihood ~ Minimizing MSE