Hypothesis Quiz Selected Answers Summer 2024

CS7641: Machine Learning

Prepared by: Theodore J. LaGrow

Last Updated: 05/29/24

The following are curated answers generated by the Teaching Staff for the article, “Robust T-Loss for Medical Image Segmentation” by Gonzalez-Jimenez et al. I have tried to highlight key details needed for each question. These answers are examples in no particular order.

Question 1

Much like Rule 6 in Ten simple rules for structuring papers by Konrad Kording and Brett Mensh, what is the gap the authors provide? Please provide 400 words or less.

- For this question, we are really looking for what the author’s claim for the gap as well as some explanation for why this is the gap. For full completeness, you need to give explanation or evidence.

Responses:

The training sets currently used for medical image segmentation are noisy and incorrectly labeled in about 5% of the data. The gap that the author’s set out to resolve is developing a new deep learning algorithm for medical image segmentation that is less sensitive to the quality of the training set compared to previous methods which include Convolutional Neural Networks (CNNs) and Visual Transformers (ViTs). In the authors’ opinion, the solution to this gap is using a modified version of robust loss functions described in the paper, named T-Loss.

In this paper, the authors address a critical gap in the field of medical image segmentation, specifically related to handling outliers and noisy labels. It sets up the context about medical image segmentation and the significance of regions of interest being affected by noise of varied levels and types. It also talks about traditional loss functions and their ability to handle outliers in a robust manner, or the lack thereof. The proposed solution is the T-Loss, a novel loss function based on the negative log-likelihood of the Student-t distribution. It is promised to be a simpler and more efficient loss function which is then backed up using experimental results with the Dice scores as the metric for segmentation accuracy. By addressing this gap, the authors contribute to improving the reliability and accuracy of medical image segmentation models.

Rule 6 of the “Ten Simple Rules” document addresses the paper’s purpose. More specifically, communicating to the reader how the new technology solves an existing problem and why the solution is important. The T-Loss paper outlines a common problem of using Convolutional Neural Networks (CNNs) and Visual Transformers (ViTs) of obtaining large amounts of training data. In medical imaging the data is costly due to labeling each pixel which requires human expertise. One possible solution is to obtain labels through automated mining or crowd-sourcing methods. This method produces data labels with a high level of noise. To correct the noisy labels several approaches have been studied such as label correction, estimated noise transition matrix, and robust loss function (the purpose of the paper). The robust loss function seems the most promising but is understudied. This paper helps fill the gap of analyzing several traditional robust loss functions and a new one, T-Loss. The paper does a good job of structuring research on robust loss by covering previous loss solutions, introducing a new loss (T-Loss), clearly outlining the datasets and metrics used the experiments to show results and discussing findings. The paper does fill the “gap” of adding research of robust loss in medical imaging and lays out a clear blueprint to continue adding the body of research on the topic.

The gap that the authors provide is the need for large amounts of annotated data when developing state-of-the-art segmentation models. The authors say that "supervised training of CNNs and ViTs requires large amounts of annotated data" and specifically for the medical domain, that "obtaining these annotations can be affected by human bias and poor inter-annotator agreement". Additionally, the authors claim that the quality of the collected datasets is difficult because there is a large amount of noise induced in the data. All of these aforementioned reasons stated by the authors contribute to the gap that the authors are trying to close with this paper.

Field gap: Large amount of annotated training data for medical image segmentation is hard to obtain and the annotations can be affected by multiple reasons causing the data with high levels of noise.
Subfield gap: Robust loss function provides a simpler solution with a single modeling component. Other approaches have limitations such as more hyper-parameters, or complex training procedures.
Gap within the subfield: Traditional robust loss functions are vulnerable to memorizing noisy labels.
Field gap: Large amount of annotated training data for medical image segmentation is hard to obtain and the annotations can be affected by multiple reasons causing the data with high levels of noise.
Subfield gap: Robust loss function provides a simpler solution with a single modeling component. Other approaches have limitations such as more hyper-parameters, or complex training procedures.
Gap within the subfield: Traditional robust loss functions are vulnerable to memorizing noisy labels.

Question 2

What is the hypothesis of the article? Please provide 100 words or less.

- For this question, you need to state the hypothesis for which the authors pose a question and their predicted outcome. Evidence will help support the hypothesis prediction.

Responses:

The article hypothesize that robust loss functions are a superior solution for semantic segmentation in medical imaging. It eliminates modifying network architecture, complex training, and multiple hyper parameters. The biggest drawback of robust loss functions has been memorizing noisy labels. The T-Loss eliminates the noisy label problem with the negative log-likelihood of the Student-t distribution. The simplest version uses a single parameter, learns label noise during backpropagation, and eliminates the Expectation Maximization steps.

The paper proposes a new kind of robust loss called T-loss based on a negative log-likelihood of the Student-t distribution. It states that this loss will introduce a single parameter during the back-propagation phase which can be trained along with the model hyperparameters to learn the tolerance level of the noise. The paper states that this will reduce complex calculations and the results on two benchmark medical datasets outperform state-of-the-art robust loss functions, especially when there is high noise. The dice scores also remain consistently high as compared with other loss functions.

The problem statement indicates that many traditional robust loss functions are vulnerable to memorizing noisy labels. The proposed solution or hypothesis that is provided is the introduction of a new loss function (T-Loss), which is based on the negative log-likelihood of the Student-t distribution. The hypothesis also mentions that T-Loss can adaptively learn an optimal tolerance level to label noise directly during backprop while eliminating the need for additional computations.

The hypothesis is that assuming error terms follow a Student-t distribution allows for significantly larger noise tolerance compared to the usual Gaussian form. Additionally, the authors hypothesize that optimizing the parameter v jointly with the model parameters w using gradient descent will enable the model to dynamically adjust its sensitivity to label noise during training. This approach is expected to improve segmentation accuracy and robustness, particularly in high-noise conditions, without requiring complex modifications to the network architecture or additional computational steps. The authors propose that the T-Loss will outperform traditional and other robust loss functions in medical image segmentation tasks.

Question 3

What specifics from article do you personally need to look up for understanding of claims? How will this information provide context to the claims? This can include background, notation, and citations. Please provide 400 words or less.

- For this question, we will give points more often than not. This is an exercise to help you go look up specific citations, math notation, or even assumed concepts. We do not expect the full word limit to be used, but we hope that the few examples you state will have some explanation to help further your understanding. The is a fine line between diving into the rabbit hole of citations and understanding the current paper.

Responses:

I needed to look up the following:

- Gain more background information about the methods currently available for medical image segmentation. I paid special attention to gain some understanding about robust loss functions since they are the major focus of the paper. Without this necessary background information, I could have made fewer comments about the paper.

- The specifics of the method used to simulate the real risk of Robust T-Loss for Medical Image Segmentation error (reference 15). Since the results presented in the paper highly depend on the simulated error, I found it important to achieve a better understanding of how and what type of errors were applied to the training set.

- Understanding nu (ν) as this is a key parameter in the T-Loss function, the dynamically optimized parameter that controls the sensitivity to outliers, likely due to error.

What noisy labelling is and common labelling techniques in order to understand the context and compare the proposed T-loss with these approaches.

The paper, Sun, J., Kabán, A., Garibaldi, J.M.: Robust mixture clustering using Pearson type VII distribution. Pattern Recognition Letters 31(16), 2447–2454 (2010) to understand why e Student-t distribution is robust to outliers.

An overview of existing robust loss functions, such as Mean Absolute Error (MAE), Reverse Cross Entropy (RCE), and Generalized Cross Entropy (GCE), and their limitations in handling noisy labels.

I was not familiar with the concept of Dice scores, and a quick internet lookup brought me upto speed about its commonness for segmentation accuracy. The major point I was not too convinced after reading the article was the following claim: By controlling its sensitivity with a single parameter, the T-Loss adapts to varying noise levels without requiring prior information about the noise distribution. While there is experimental proof in the form of graphs suggesting the same, I was hoping to have some more information about how and why a single parameter suffices the task at hand. Providing this information will make a strong case for T-Loss to be one of the SOTA loss functions in the near future.

Methods were more or less clear, however, certain key terms (e.g. t-loss and many of the losses under 3.2 Setup, gamma normalization, nnU-Net) warranted a look up and/or refresher to grasp at what I deemed to be an adequate level, as the authors appeared to assume prior knowledge of many of the glossed over terms, methods, treatments, etc. Additionally, statistical terms, tests and metrics such as Student-t distribution, Tukey's HSD and dice loss and score and different activation-function-dependent and very case-specific probability functions warranted a Google search or a sit and think to fully grasp.

There's an initial need to better understand what is the label correction method, and what is the loss function correction method since the two methods were presented as suboptimal to the robust loss function due to their complex nature. When examining reference 28 (Learning from Massive Noisy Labeled Data for Image Classification) the paper clearly shows the complexity of the adjustment needed to the machine learning model to correct the labels prior to learning. Further, it is important to point out that the data used in the paper was clothes images, which don't require much detailed annotation. Another topic that needs a deeper review is the different morphological transformations used to mimic human error. The authors have used erosion, dilation, and affine transformations to simulate the noise in the segmented data due to human error, but it is unclear if these transformations truly represent human error. The author uses the same strategy in reference 15 (Superpixel-Guided Iterative Learning from Noisy Labels for Medical Image Segmentation), which in turn reference three other papers that don't offer any rational for the correctness of the induced noise as it resembles the human error. Lastly, math review was needed to understand the properties of the Student's t-distribution. This search is important to grasp the fundamentals of the t-loss function. The review shows that the key propriety of the Student's t-distribution is the variable degree of freedom parameter that controls the sensitivity to outliers, which in turn can be dynamically adjusted based on the level of noise, which the entire premise of the paper.

Following articles are helpful to understand the claims authors made in the paper. Especially the first article is more beneficial as it gives a general overview of the robust loss function, which is the central point of the article.

[1] Barron, J. T. (2019). A general and adaptive robust loss function. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4331-4339).

[2] Forbes, F., & Wraith, D. (2014). A new family of multivariate heavy-tailed distributions with variable marginal amounts of tailweight: application to robust clustering. Statistics and computing, 24(6), 971-984.

[3] Liu, S., Liu, K., Zhu, W., Shen, Y., & Fernandez-Granda, C. (2022). Adaptive early-learning correction for segmentation from noisy annotations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2606-2616).

[4] Ma, X., Huang, H., Wang, Y., Romano, S., Erfani, S., & Bailey, J. (2020, November). Normalized loss functions for deep learning with noisy labels. In International conference on machine learning (pp. 6543-6553). PMLR.

While I was familiar with the PDF of the Student-T distribution , I had not seen its form in the D-dimensional setting. For me, I believe understanding the formulating of (1)-(4) are quite important details for implementation. Additionally, I personally needed to look up the notion of power law to understand how it makes the overall loss function more "robust".

Question 4

After reviewing and re-reviewing the article, what limitation did do discern? These can be simple or more complex observations. Please provide reasoning in 400 words or less.

- This is the heart of the assignment. To be critical while reading is an important skill to develop. Even though this article is highly recognized at a top conference, there are always connotations and limitations to the paper. Science and engineering cannot be done in a vacuum. For these answers, we are looking for reasoning behind the limitations you present. You want to be concise rather than harsh and overbearing. Below are some examples from the Teaching Staff.

Responses:

Claiming that there are no public benchmarks with real noisy and clean segmentation masks, the authors artificially inject additional mask noise in the two datasets to test the model's robustness to low annotation quality - ISIC 2017 dataset (2000 images), Shenzhen (296 images). The synthetic noise injection could be structured to follow the PDF that is discernable by the T loss and perform well under that expected distribution of noise.

As for the shortcomings of this paper, the first issue is that the training was conducted for only 10 epochs, which may not be sufficient for the model to converge. If the model has not fully converged, the results may be unreliable and potentially meaningless. Adequate training duration is crucial for obtaining valid and robust outcomes.

Another limitation is that the noise was manually injected into the dataset, while the original dataset already contains some level of noise. This approach means the paper is essentially comparing the performance on an already noisy dataset with an artificially more noisy version, rather than evaluating the robustness of the model against naturally occurring noise. This may not accurately reflect real-world conditions where the nature and extent of noise can vary significantly.

Additionally, the comparisons in the paper are exclusively made with the U-Net model. While U-Net is a well-known and widely used architecture, it is not the latest or best-performing model as of 2023, the publication year of the paper. The authors should have considered using the latest state-of-the-art models for a more comprehensive and up-to-date evaluation of the T-Loss's performance.

Moreover, the simplification of the covariance matrix to the identity matrix in the T-Loss derivation could overlook important dependencies between pixel annotations in medical images. This assumption, while computationally convenient, might limit the effectiveness of the loss function in capturing the true complexity of the data.

Furthermore, the experiments rely on artificially noisy data, and the paper does not explore the impact of different initializations or optimization strategies for the single adaptive parameter v. A deeper analysis of parameter sensitivity and its effects on various noise types would provide a more robust understanding of the T-Loss's reliability.

Lastly, the scope of the application is limited to two specific tasks: skin lesion and lung segmentation. The generalizability of the T-Loss to other medical imaging tasks, such as brain MRI or retinal imaging,

remains untested. Extending the evaluation to a broader range of datasets would help in assessing the overall applicability and robustness of the proposed loss function.

Major limitations include:

The small size of both datasets used to compare the performance of different robust loss functions.

Use of only 3 random seeds for each algorithm configuration.

Minor limitations include:

- A widespread lack of important information in section 3.1 and 3.2 which are the Methods section. I needed to read the referenced paper to understand how and what type error was applied to the images in the training set and I think the authors could have included an explanation in their manuscript to facilitate the reader's understanding. I am generally skeptical when technical details are left out as this seems an attempt to prevent expert readers from evaluating the technical quality of the work, which of course affects the relevance of the results.

- The authors do not provide details about how they optimized T-Loss function hyperparameters nor about hyperparameters optimization for the other functions used for comparison. Hyperparamters optimization affects performance and if the authors were not thorough in optimizing them at best for each function examined, this could artificially skew the results in favor of one method over the other.

- The exclusive focus on the dice score to evaluate performance. If I were to convince the readers to adopt the amazing new method developed, I would compare it to a wider range of existing and competing methods and I would examine more aspects of its performance and computation time.

- In the Conclusion section, the authors mention that at least one other method (15) performs as well as, if not even better than, their T-Loss on one of the datasets used, the ISIC dataset. It is dubious why the authors have not included this method in the set of alternative methods used for comparison and if they have omitted in their examination even other methods of similar or better performance compared to T-Loss.

The effectiveness of the T-Loss is evaluated only on simulated label noise, which may not fully represent the complexities of real-world noisy annotations.

The experiments are limited to two medical image segmentation datasets, and the generalization of the T-Loss to other domains or modalities is not explored.

The computational cost and convergence properties of the T-Loss, particularly with respect to the dynamic optimization of the v parameter, are not discussed in detail.

The authors do not provide a comprehensive comparison with other state-of-the-art methods for handling noisy labels in segmentation tasks.

---

1. The results given in the paper are from experiments performed on two datasets. It would be good to conduct experiments and share results on more datasets to ensure generalizability of the results.

2. The two datasets used for the experiments pertain to skin lesion and lung segmentation. It would be good to experiment on other types of medical datasets as well to ensure generalizability.

3. Due to absence of public benchmark with real noisy data, noise has been artificially injected into the two datasets. Even though this simulation is claimed to be like the real presence of noise, it would be good to try and find datasets with real noise to verify the results.

The label noise was artificially injected and there is no public benchmark for real-world noise. Real-world medical image annotations may have noise characteristics that differ significantly from those simulated in the experiments. The T-Loss function is based on the negative log-likelihood of the Student-t distribution, which is assumed to effectively model the noise in the annotations. This assumption may not hold for all types of label noise encountered in practice.

In my opinion, it would have been helpful to fully outlay their methodology in the methods section. It seemed tidbits of procedural discussion that ought to have been in the methods section were distributed throughout the paper. Additionally, the authors seemed to provide more context/rationale for some decisions than others. Discussion of parameter tuning, approach and effects of the ablation study and other such specifics may have given the reader a broader sense of the scope of their work. Discussion of future work, directions and preliminary sense for how generalizable the authors' methods and contribution may be were missing.

I think a limitation of this article, as is with any novel, state-of-the-art approach, is how well it scales and translates to different use cases. Here, they test two types of datasets, one on skin cancer and one on chest radiograph images, which really only cover a very small subset of the potential applications in the medical field. While this is a good start, I would be interested in seeing how this application scales to, say, images multiple orders of magnitude smaller or larger in size. Additionally, more specific to this paper, the authors acknowledge that there is no public benchmark with real noisy and clean segmentation masks; thus, they manually add noise to each of the two datasets to simulate low annotation quality. How well does this simulation reflect actual annotation uncertainty in the real world? Nuanced differences here may have a drastic effect on the usability of this solution, depending on the accuracy requirements of the use case.

Question 5

What might be next steps to further this research? There is no correct answer here, however your answer must be grounded in reasoning.

Please then provide a hypothesis on what you'd like to investigate based on the knowledge gained in this exercise. Be concise with detail.

Limit full response to 400 words.

- This question is looking for a hypothesis for further research you'd conduct after reading the article. What sparks your interest? Having evidence or a direction will help support your hypothesis.

Responses:

First, it is essential to conduct experiments with longer training durations to ensure the model reaches full convergence. This would validate the robustness of the T-Loss over extended training periods and provide more reliable results. Testing the T-Loss on naturally noisy datasets without artificially injected noise is another crucial step. This would better reflect real-world conditions and validate the robustness of the T-Loss in practical applications. Evaluating the performance of the T-Loss using the latest state-of-the-art models, such as transformers and more advanced CNN architectures, would provide a comprehensive understanding of its effectiveness compared to the most recent advancements in medical image segmentation. Conducting a thorough analysis of the sensitivity of the parameter v and its initialization is also necessary. Exploring different optimization strategies and their impact on performance would provide deeper insights into the robustness of the T-Loss. Extending the evaluation of the T-Loss to a wider range of medical imaging tasks, such as brain MRI, retinal imaging, and other modalities, would test the generalizability of the T-Loss across different types of medical data. Integrating domain-specific priors or constraints into the T-Loss function could improve segmentation accuracy and robustness. For example, leveraging anatomical knowledge in medical images could enhance model performance. Finally Investigating the combination of T-Loss with other techniques like data

augmentation and semi-supervised learning methods could further enhance the robustness and performance of segmentation models.

To test this hypothesis, the following steps could be undertaken. First, modify the T-Loss to include anatomical priors relevant to specific medical imaging tasks being investigated. For example, incorporating structural information about organs or tissues can provide additional guidance to the segmentation model. Next, apply sophisticated data augmentation techniques that mimic real-world noise patterns and variations more closely. This could include geometric transformations, intensity variations, and synthetic noise that resembles actual medical imaging artifacts. Conduct extensive experiments on diverse and real-world noisy medical datasets, such as those from public medical imaging challenges, to compare the performance of the enhanced T-Loss against both the original version and other robust loss functions using state of the art models. Finally, perform detailed statistical analyses to evaluate the significance of the improvements observed. This includes metrics such as dice score, precision, recall, and robustness under varying noise levels.

Future work is mentioned in the paper itself - The trade-off in terms of performance, computational cost, and ease of adaption to different scenarios remains to be investigated. Similarly, combinations of the T-Loss with superpixels and/or iterative label refinement procedures are still to be explored.

The 1st point can be performed by a more extensive research and experiments yielding substantial results for comparison in the performance, cost, and ease of adaption in other various scenarios. The latter point is a more novel scenario and an hyper-extension of the current research.

Superpixels are compact, perceptually meaningful image regions. Combining them with the T-Loss can enhance segmentation accuracy. Research can be in the direction of what superpixels are, their key properties, and how we can amalgamate it with T-Loss.

After initial segmentation, iteratively refining labels using additional information and ML/DL algorithms will possibly reduce the memorization of noise labels even more and improve convergence after passing it through the training iterations.

Other future works could be trying to improve on the limitations mentioned in the previous 2 questions as well as exploring with different datasets, evaluation metrics, and combine T-Loss with other loss functions to benchmark against SOTA methods.

The only method of performance evaluation presented in the paper is comparison of T-loss with other loss functions. Evaluating other performance metrics, evaluating computational cost, ease of adaptation to different scenarios etc. can be some good next steps to further this research. Experimenting on a wider variety of datasets, with real noise data would also be good.

Since the existing research was done using synthetic data which is as close as possible to the real noisy data, if tested on more diverse datasets, I hypothesize that T-Loss will exhibit promising results and will outperform traditional loss functions, like how it performed on the synthetic datasets. I also hypothesize

that the computational costs will not significantly increase rise because of its simple formulation using a single parameter that can adaptively learn an optimal tolerance level to label noise directly during backpropagation, eliminating the need for additional computations such as the Expectation Maximization (EM) steps.

Some next steps could include:

- Validating T-Loss on diverse datasets of medical imaging. This may include MRI and CT scan imaging on various anatomical regions.

- Developing a real-world noise benchmark in medical imaging. This would allow us to check the noise assumptions T-Loss makes when modeling noise with a the negative log likelihood of Student's T distribution.

- Investigate the adaptation of T-Loss to multi-class segmentation. This may involve some type of multivariate Student T distributions.

- Analyze the computational cost and efficiency of T-Loss, particularly in large-scale or real-time applications. Optimizing the implementation and exploring hardware accelerations (e.g., GPU optimizations) can help make T-Loss more practical for clinical use.

- Perhaps exploring alternative noise models beyond Student's T and Gaussians. There may be other distributions that better capture characteristics of real-world label noise in medical imaging.

A hypothesis that I would like to investigate may include some of the points detailed above and could read like this:

"The T-Loss function, when adapted for multi-class segmentation and validated on a broader range of medical imaging datasets with real-world noise characteristics, will demonstrate superior robustness and generalizability compared to existing robust loss functions, leading to improved segmentation accuracy in diverse clinical settings."

This hypothesis touches on dataset diversity, multi-class adaptation, and utilization of real-world noise.

Version Control

Revision	Date	By	Version	Change Description
1	05/24/24	TJL	0.1	• Initialization of template. • Compile answers and synthesize information.
2	05/29/24	TJL	1.0	• Proofreading. • Posting to Canvas.