Global Response

Dear Reviewers:

Thank you for your efforts on our work, for the constructive suggestion and insightful comments on our work, and for the affirmation of our work. Here, we succinctly summarize the strengths of our work highlighted by reviewers and the additional experimental enhancements we have implemented.

\*\*Strengths highlighted by reviewers \*\*

1.The proposed theoretical analysis framework is valuable, reasonable, interesting (Reviewer UDgG , Reviewer o8gz, Reviewer YKuS); could potentially generate further insights (Reviewer EJUe), deserves further exploration (Reviewer UDgG), and serves as a good stepping stone for future follow up works (Reviewer EJUe).

2.CRReLU, the derived activation function from the theoretical framework, outperforms baselines on image classification tasks with ViT variants as well as LLM fine-tuning (Reviewer UDgG, Reviewer o8gz, Reviewer EJUe, Reviewer YKuS).

3.The presentation and theoretical analysis are easy to follow; the paper is written clearly and concisely, and is easy to read. (Reviewer YkuS, Reviewer EJUe).

\*\*Additional Experimental Enhancements\*\*

1. \*\*Stronger training recipe on Transformer-based structures.\*\*

We perform \*\*300 epochs\*\* training to evaluate the performance of GELU, ELU, PReLU, StarReLU [1], SiLU(Swish), Mish and CRReLU(ours) activation functions on ViT-Tiny (training curves plotted in Figure1 and Figure2), DeiT-Tiny (training curves plotted in Figure3 and Figure4) and DeiT-Base (training curves plotted in Figure5 and Figure6).

2.\*\* Standard training recipe on convolution networks\*\*

We further perform \*\*300 epochs\*\* training to test the performance of GELU, ELU, PReLU, StarReLU (ReLU for ResNet), SiLU(Swish), Mish and CRReLU(ours) on ResNet-18 (training curves plotted in Figure7 and Figure8), ConvNeXt-Tiny (training curves plotted in Figure9 and Figure10) and ConvNeXt-Large (training curves plotted in Figure11 and Figure12).

Furthermore, in order to further explore the learned $\epsilon$, we visualize variations of the learned $\epsilon$ of DeiT-Base in Figure13 (CIFAR10) and Figure14 (CIFAR100) with 12 layers respectively; and we also visualize variations of the learned $\epsilon$ of ConvNeXt-Large in Figure15 (CIFAR10) and Figure16 (CIFAR100) with 36 blocks respectively. All the experiments are conducted under exactly the same conditions: same optimizer, same scheduler, same learning rate, same data augmentation, same random seed, same hardware (4\*RTX3090), and so on.

[1] MetaFormer Baselines for Vision. TPAMI, 2024.

Dear Reviewer YkuS:

Thank you for your efforts on our work, for your affirmation of our work and for your constructive suggestion. Based on your insightful comments, we summarize the strength of our work as follows:

1. The proposed theoretical analysis tools are \*\* interesting\*\* for activation functions in deep learning; the designed activation function is \*\*new\*\*.

2. The presentation and theoretical analysis are \*\* not hard to follow\*\*, and the structure is \*\* clear\*\*.

\*\*Response to W1\*\*

Thank you once more for your insightful comment and question. This paper primarily focuses on bridging the theoretical gap in understanding activation functions. We place a greater emphasis on the theoretical analysis in the paper, with the same deliberate intention of conveying this concept to researchers and readers. We tend to view the derived activation function as an example or a preliminary exploration, akin to the SquareReLU for the StarReLU [1]. As you noted, the StarReLU activation function is characterized by two learnable parameters, thereby enhancing its flexibility and diversity in neural network architectures; and in the work [1], as mentioned, also showcases its superior performance on the MetaFormer, demonstrating the effectiveness and versatility of the StarReLU. This, however, does not imply that CRReLU lacks novelty. The two ReLU variants were proposed based on different motivations, each embodying its own innovation: StarReLU exhibits its creativity in the design of activation functions grounded in practical insights, whereas CRReLU showcases its innovation rooted in theoretical foundations. We believe that future research will explore the application of CRReLU in a manner analogous to the development of StarReLU from SquareReLU. To further validate the effectiveness of CRReLU, we compare its performance to that of StarReLU, and using StarReLU as one of the baselines. The results demonstrate that CRReLU achieves performance comparable to StarReLU. We have conducted a more in-depth analysis of StarReLU; however, the two learnable parameters, which grant its greater flexibility and complexity, also render it challenging to the analytical solution of the entropy equations; hence making it difficult for us to provide a thorough theoretical explanation currently for why StarReLU is a superior version of ReLU activation. We intend to expand on a discussion in the “Limitation” section, with the hope that future research may build upon these two versions of ReLU activation to drive further improvements.

\*\*Response to W2\*\*

Thank you once more for your constructive suggestion. In the experimental section, we originally hoped that researchers and readers would focus more on the EAFO framework presented in this paper. And following your constructive suggestions, we would like to strengthen the experiments accordingly. We perform 300 epochs training with ViT-Tiny, DeiT-Tiny, DeiT-Base, ResNet-18, ConvNeXt-Tiny and ConvNeXt-Large; and add StarReLU as one of the baselines on ViT/DeiT and ConvNeXt. Furthermore, we provide the visualization of test loss and test accuracy over the course of training, which can be found in the Global Response.

\*\*Response to W3\*\*

Thank you once more for your constructive suggestion and comment. We will provide a more systematic and comprehensive discussion on the existing activation functions in the appendix according to the latest survey[2], titled "Further Discussion on Existing Activation Functions". Our initial intent on the bold in Table 3 (GELU and CRReLU in the row of DeiT-T) was to convey that the results are comparable. However, such might have caused confusion; to clarify, we would remove the bolding from the results that are not optimal. The space before the comma in line 248 will be removed, and commas will be added to separate the items in Equations (8) and (9). Visualization of training iterations for the convergence speed and visualization of the learned $\epsilon$ are available in the Global Response.

\*\*Response to Q1\*\*

Thank you once more for your constructive suggestion and insightful question. We have visualized variations of the learned $\epsilon$ of DeiT-Base and ConvNeXt-Large in the Global Response.

\*\*Response to Limitations\*\*

Thank you once more for your constructive comment. The StarReLU activation function, which incorporates scaling and bias operations, has demonstrated promising results in the context of MetaFormer. We intend to expand on the discussion in the "Limitation" section, with the hope that future research may build upon these two versions (both practical insights and theoretical foundations) of ReLU activation to drive further improvements.

Finally, thank you once more for your efforts on our work, for your affirmation of our work and for your constructive suggestion.

Warm regards,

Authors of submission 15900

[1] MetaFormer Baselines for Vision. TPAMI, 2024.

[2] Three Decades of Activations: A Comprehensive Survey of 400 Activation Functions for Neural Networks. ArXiv, 2024.

Dear Reviewer o8gz:

Thank you for your efforts on our work, for your affirmation of our work and for your constructive suggestions. Based on your insightful comments, we summarize the strength of our work as follows:

1. The theory is \*\*reasonable and interesting\*\*.

2. A \*\*significant accuracy\*\* has been gained.

\*\*Response to Weaknesses and Questions \*\*

Thank you once more for your constructive suggestions. Our initial intention is indeed to conduct a fair comparison; therefore, we trained the model for 100 epochs. Subsequently, following your constructive suggestion, we expand the training with a stronger training recipe and train for 300 epochs. Furthermore, we also conduct corresponding experiments on convolutional networks, including ResNet-18, ConvNeXt-Tiny, and ConvNeXt-Large for 300 epochs. The training curves, encompassing test accuracy and test loss, are visualized in the Global Response. The results indicate that under a stronger training recipe, the superiority of CRReLU still holds, especially the convergence speed. We also further visualize the activation decision boundary, but we find that: after 300 epochs of training, the accuracy gap between different activations has become relatively small, and when there are many classification categories, it is difficult to intuitively and clearly explain that a particular activation's decision boundary is significantly better than others. If we select only two or a few classes for visualization, although the results are intuitive, they also lack representativeness, so we discard the presentation of visualization results for the activation decision boundary. While, to delve deeper into the understanding of learned $\epsilon$, we visualize the changes of the learned $\epsilon$ of DeiT-Base and ConvNeXt-Large, which are presented in the Global Response.

Finally, thank you once more for your efforts on our work, for your affirmation of our work and for your constructive suggestions.

Warm regards,

Authors of submission 15900

Dear Reviewer EJUe:

Thank you for your efforts on our work, for your affirmation of our work and for your constructive suggestions. Based on your insightful comments, we summarize the strength of our work as follows:

1. The paper is \*\*written clearly and concisely\*\*, and is \*\*easy to read\*\*.

2. The theoretic framework is a \*\*great\*\* approach, could \*\*potentially generate insights\*\*; and serves as \*\*a good stepping stone\*\* for future follow up works.

\*\*Response to Weakness 1\*\*

Thank you once more for your insightful comments. As we mentioned in Line119~121 of the paper, “in binary classification, the feature extraction layer can be conceptualized as transforming the shape of mixture distribution, thereby enabling the final fully connected layer to separate two distributions with a hyperplane.” There is often the final fully connected layer behind the feature extraction layer; hence, the better separability of features would be mapped by the final fully connected layer (which can be regarded as a hyperplane) to better classification space. In other words, better separability of features results in greater probability of a better classification performance, and the probability comes from the final fully connected layer (which is easy to train and can be trained approaching to perfect); and we believe that with sufficient training, such probability would be quite high. The insightful point you mentioned that the entropy in continuous random variables would change with the scale indeed might not have an impact on the classification performance under the current static design of activation functions in MLPs. While, we also note that if such framework is generalized to dynamic activation optimization or network structures with changing scales, such as KANs, your viewpoint is highly valuable, in which regularizing the entropy function with respect to the scale might be necessary. We would add a discussion on this point in the paper to alert future researchers to pay attention to this issue.

\*\*Response to Weakness 2\*\*

Thank you once more for your constructive suggestions. Following your suggestion, we conduct additional experiments with CNN-based architectures. We visualize their training curves, along with other supplementary experiments, in the Global Response.

\*\*Response to Weakness 3\*\*

Thank you once more for your insightful comments. Each experimental result presented is based on one run of training and all the experiments are conducted under exactly the same conditions. We did not provide error bars mainly for two reasons: firstly, from the perspective of paper layout and aesthetics, providing error bars for every piece of data can result in the data appearing too cluttered; secondly, we expect that readers of this study will emphasize more on the theoretical analysis process over the experimental results.

\*\*Response to Question1\*\*

Thank you once more for your insightful comments. We have selected the baselines to some extent based on a previous survey[1]. The survey is mainly composed of the five parts; and in each part, we chose one or two of the most classic activation functions as our baseline: we chose SiLU (Swish) in "Logistic Sigmoid and Tanh Based AFs"; we chose ReLU (now some for the StarReLU) in "Rectified Activation Functions"; we chose ELU in "Exponential Activation Functions"; we chose PReLU in "Learning/Adaptive Activation Functions"; in "Miscellaneous Activation Functions", we chose Mish for the first part "Softplus Activation Functions" and GELU for the second part "Probabilistic Activation Functions".

\*\*Response to Question2\*\*

Thank you once more for your insightful comments. We select ReLU as our starting point partly due to its advantageous characteristics, including the properties of ReLU, its derivative and second-order derivative: Firstly, with $y''=0$ almost everywhere, the first term of $\eta(x)$ in Equation (7) becomes 0; Secondly, with $y'=1(x>0)$, it leads to a straightforward expression that $\eta(x)=p'(x)$. Such attributes contribute to a relatively concise derivation. Additionally, we chose the ReLU activation function with the purpose of exploring the framework application to the extensive family of non-invertible activation functions, which is expected to lay a foundation for future research. We have also tried starting with Mish, Swish and GELU, but their analytical forms are too complex, and we have not been able to derive a strict analytical approximation so far. While, we have tried to add the correction term derived from ReLU to them, resulting in the corresponding CRMish, CRSwish, and CRGELU. In experiments, we observed that the newly obtained activation functions do perform better than the original activation functions, which appears to stem from their approximation to the ReLU in the presence of sufficiently large positive or negative input values. However, we hope to conduct further investigation and derivation rather than jumping to hasty conclusions, for the sake that we have not only seeking practical efficiency but also pursuing theoretical rigor in our pursuit of excellence and scientific research.

\*\*Response to Question3\*\*

Thank you once more for your insightful question. We believe that in the future, there would be more studies to validate the practicality of this framework. For now, it's difficult for us to confidently answer your visionary question. Meanwhile, we hope that this framework will not only play a role in the design of static activation functions but also be applied to dynamic activation optimization and more novel network structures (such as KANs).

Finally, thank you once more for your efforts on our work, for your affirmation of our work and for your constructive suggestions.

Warm regards,

Authors of submission 15900

[1] Dubey, S. R., Singh, S. K., & Chaudhuri, B. B. (2022). Activation functions in deep learning: A comprehensive survey and benchmark. Neurocomputing, 503, 92-108.

Dear Reviewer UDgG:

Thank you for your efforts on our work, for your affirmation of our work and for your constructive suggestions. Based on your insightful comments, we summarize the strength of our work as follows:

1. The insights into the problem of activation function optimization and the analytic structure of improved activations is \*\* valuable\*\* and \*\*deserves further exploration\*\*.

2. The \*\*novel\*\* activation function \*\*outperforms\*\* common baselines.

\*\*Response to Weakness1\*\*

Thank you once more for your insightful and visionary comments. We simplify the analytic study with the idea that better classification performance implies better activation function. As you commented, such idea may not be quite applicable to several certain scenarios, such as diffusion process. Despite this, it can still find applicability in numerous scenarios that can be framed as the generalization of classification problems, such as object detection / image segmentation (pixel-level classification), large language models reference (word-level embedding classification), deep reinforcement learning (classified action selection), and so on. We believe that this framework is potentially applicable in the similar scenarios. Furthermore, we consider that the method would not be used as the only method to identify improved activation functions, but should be one.

\*\*Response to Weakness2 and 3\*\*

Thank you once more for your constructive suggestion. Following your suggestion, we conduct additional experiments with CNN-based architectures: ResNet-18, ConvNeXt-Tiny and ConvNeXt-Large. We visualize their training curves, along with other supplementary experiments, in the Global Response.

\*\*Response to Weakness4\*\*

Thank you once more for your insightful comments. Within the scope of our current framework, specifically in the design of static activation functions in MLPs (such as CRReLU), the fact that each activation function results in a different set of optimized network weights might not affect the final outcome. This is because we are optimizing the network together with the activation function as an integral part of it (including the hyperparameter $\epsilon$ in CRReLU, which is also placed in the optimizer). However, when it comes to the consideration of applying the framework to dynamically optimizing activation, we believe your insight is of great importance. Actually, your insightful comment points out one of the challenges we encounter in handling dynamic activation optimization. When aiming to dynamically optimizing activation during iterative training, we must account for its effect on the entire network, substantially elevating the algorithm’s complexity, leading to the algorithm lacking the requisite practicality.

\*\*Response to Question1\*\*

Thank you once more for your insightful question. The training pipelines for the LLM fine-tuning task are based on the standard pipelines presented in the original DPO paper. And the training pipelines for the image classification task are not the standard pipelines from the original papers. We notice that different models’ pipelines vary in their choices of optimizer, scheduler, learning rate, etc. Therefore, for a fair comparison, we standardize the training process under exactly identical settings: same optimizer, same scheduler, same learning rate, same training epoch, same data augmentation, same random seed, same hardware environment, and so on. Furthermore, we conduct experiments with a stronger training recipe, where we train on ViT-Tiny, DeiT-Tiny, DeiT-Base, ResNet-18, ConvNeXt-Tiny and ConvNeXt-Large for 300 epochs, while maintaining all other conditions constant. We present the visualization of test loss and test accuracy over the course of training, detailed in the Global Response.

\*\*Response to Question2\*\*

Thank you once more for your careful review of our work. We will correct the typo. We further conduct a comprehensive review of the entire paper and do not identify additional typo errors.

Finally, thank you once more for your efforts on our work, for your affirmation of our work and for your constructive suggestions.

Warm regards,

Authors of submission 15900