Supplementary Material

Hailun Liang[1(✉)], Sunyuan Qiang[1], Hui Ma[1], Jun Wan[1,2,3], and Yanyan Liang[1]

[1] Macau University of Science and Technology, Macau SAR, China
helenzehai@gmail.com
[2] MAIS, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[3] SAI, University of Chinese Academy of Sciences, Beijing, China

1 Experiments Setting

Datasets The Cambridge-driving Labeled Video Database (CamVid) [1] is a popular dataset in computer vision research, created for semantic segmentation tasks. It features car-driving videos from Cambridge, meticulously annotated at the pixel level with 32 classes like roads, buildings, and pedestrians. We inherit a modified version of the 11 classes of the Camvid dataset. CamVid is known for its high resolution and intricate scenes, making it essential for testing image understanding algorithms. We assess SC-CoreSet using the CamVid dataset, following the PidNet configuration. The specific settings are detailed in the table provided. Initially, 10% of the training data forms the annotated dataset $\mathcal{D}^a$, with subsequent iterative sampling of 5% new data $\mathcal{D}^s$ from the remaining training set, constituting the unlabeled data pool $\mathcal{D}^u$. Due to high similarities in street scenes, an initial subset is randomly chosen from $\mathcal{D}^u$, from which m samples are queried.

Table 1. Evaluation datasets.

Dataset	Classes	Train	Valid	Test	Initial labeled	Budget	Image Size
CamVid	11	370	104	234	40	20	960×720

Implementation details We reproduce the experiments in the latest baseline PIDNet code to make the experimental results comparable at present. In order to compare with the results of DEAL, we also the Deeplabv3+ architecture with Mobilenetv2 and ResNet50 backbone. The training process involves images of size 960x720 pixels, a batch size of 4, random shuffling, training over 200 epochs, an initial learning rate of 0.005, weight decay set at 0.0005, utilizing the SGD optimizer with a momentum of 0.9. Nesterov acceleration is not employed, while horizontal flipping and multi-scale training are activated. Additionally, pixels labeled as 255 are disregarded during the training process.

2 Disscusion for Flooding Level and Pareto Optimality

The combination of flood regularization and Elastic Weight Consolidation (EWC) [2] aims to enhance the generalization capability of active learning. Flood regular-

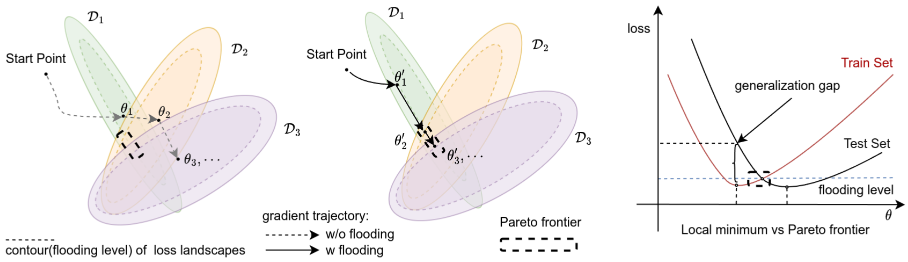

Fig. 1. Visualization of Combined Regularization Active learning samples high-uncertainty instances in each iteration. Incorporating these samples into the training set modifies the local minima in the training loss landscape. The figure's elliptical regions with colored backgrounds depict local minima and their surrounding loss basins. Discrepancies may occur between the local minima in the test set and those in the training set, which can widen the generalization gap when converging to the local minima of the training set, as illustrated in the third figure. For a one-dimensional learnable parameter, the Pareto points for the training and test sets correspond to the intersection of their loss landscapes. Stopping gradient descent at this intersection minimizes the generalization gap. Flood regularization in gradient descent involves searching along the contours of the loss landscape. When the contour value aligns with the loss at the Pareto point, the likelihood of discovering a solution with a smaller generalization gap increases. Moreover, active learning uses incremental sampling, leading to overlapping loss basins of the training set at each iteration. Constrained search ranges help mitigate the generalization gap between iterations, with the Pareto optimal regions of each training step representing the best search paths. Utilizing EWC regularization aids in identifying the shortest gradient search path.

ization is applied when there is a discrepancy between the training and target set. This discrepancy manifests as an inconsistency in the local minima of the loss landscapes for the training and test sets. Achieving a local minimum on the training set indicates a significant generalization performance gap for the test set. We shift our optimization objective from local minima to Pareto optimality[3], meaning we seek solutions that lie at the intersection of equal loss levels in the loss landscapes of both the training and target sets, as illustrated in Figure 1. At this point, we cannot improve the fitting performance of either dataset without degrading the performance of the other, which is the definition of a Pareto point. However, since optimizing using the training set does not provide visibility into the loss landscape of the target set, direct optimization may lead the gradient descent to converge at a local minimum of the training set. Therefore, we employ flood regularization to enforce that the gradient search on the training set halts at a certain level of loss contours, thereby preventing gradient descent from reaching a local minimum on the training set and potentially allowing the search process to terminate at a Pareto point.

The active learning sampling process is an incremental data acquisition process, necessitating a balance between learning from new and existing datasets. Elastic Weight Consolidation (EWC) regularization estimates the importance of model parameters about the currently learned dataset, ensuring that these crucial parameters undergo minimal changes [5]. In contrast, we update less important parameters to incorporate information from newly sampled data. Inspired by the PAC generalization loss principle, applying EWC constraints at each sampling stage helps minimize the variance between solutions, thereby enhancing the model's generalization performance [7]. Additionally, the active learning sampling process involves a continuously evolving loss basin for the training set in the solution space, implying that the local minima searched at each iteration may differ. As previously discussed, flood regularization allows the gradient descent process to halt at Pareto points corresponding to the losses of datasets from multiple incremental sampling steps.

3 Disscusion for Flooding Level Searching

In our work, we aim for the flooding hierarchy ($|\ell - b| + b$) to be consistent with the Pareto loss of both the training and testing sets. Achieving this necessitates an awareness of the geometric properties of the loss landscape in the solution space. However, since we cannot directly observe the testing set during training, the flooding hierarchy can assist in approximating the dataset. We typically sample the auxiliary dataset from cross-validation folds of the training and validation set. Additionally, estimating the possible Pareto loss relies on an auxiliary model (AM); for example, in AdaFlood [8], the authors utilize an AM trained on a cross-validated training set to estimate the loss.

We propose the following methods for estimating the flooding loss level. First, we suggest an early stopping and validation set loss combination strategy (ES-VL). The early stopping strategy [6] halts training when the model determines that the validation set $\mathcal{D}^{train}$ loss and the testing set $\mathcal{D}^{val}$ loss are no longer decreasing. This moment represents the Pareto point in the solution space between the training and validation sets. At this point, the training set loss represents the flooding level $b = \ell^{\theta^{\alpha}}(\mathcal{D}^{train})_{\text{early stop}}$. Second, we recommend using an auxiliary model (AM), for instance, by employing a pre-trained model θ^{train} and then fine-tuning it with the validation set to obtain θ^{val}. We can then verify the losses for both the training and validation sets using $\theta^{\alpha} = \alpha\theta^{train} + (1-\alpha)\theta^{val}$. The optimization goal becomes $\alpha = min_{\alpha}(\ell^{\theta^{\alpha}}(\mathcal{D}^{train}) + \ell^{\theta^{\alpha}}(\mathcal{D}^{val}))$. Then, flooding level $b = \ell^{\theta^{\alpha}}(\mathcal{D}^{train})$. The fine-tuning process minimizes linear connectivity fences in training to ensure the maximum possible connectivity between solutions of the training and validation sets. In the active learning process, continuous sampling of the training set allows us to iteratively perform connectivity fine-tuning[4] with the validation set to obtain the flooding levels at each stage, denoted as $\{b_1, \ldots, b_n\}$. Furthermore, we can experimentally explore the setting of flooding levels using Exponential Moving Average (EMA) smoothing, defined as $b_i' = \beta b_i + (1-\beta) b_{i-1}$.

The hidden variable influencing the setting of the flooding hierarchy is the model structure, which directly determines the number of parameters and consequently affects the solution space. We believe the specific value of the flooding hierarchy is directly related to the properties of the solution space, such as the scale of learnable parameters and the topological logic of neural networks. Calculating an appropriate flooding hierarchy based directly on the properties of the solution space remains a challenge. Therefore, employing two strategies, ES-VL and AM, for estimating the flooding hierarchy may expedite the search for a suitable value. However, the lack of overall awareness of the geometric nature of the loss landscape in the solution space means that manual adjustments and empirical searches for the flooding hierarchy can still prove effective.

References

1. Brostow, G. J., Fauqueur, J., Cipolla, R.: Semantic object classes in video: A high-definition ground truth database. Pattern recognition letters, 30(2), 88-97. (2009)
2. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., ... & Hadsell, R. : Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences, 114(13), 3521-3526. (2017)
3. Lin, X., Zhen, H. L., Li, Z., Zhang, Q. F., & Kwong, S. : Pareto multi-task learning. Advances in neural information processing systems, 32, (2019).
4. Juneja, J., Bansal, R., Cho, K., Sedoc, J., & Saphra, N.: Linear connectivity reveals generalization strategies. arXiv preprint arXiv:2205.12411, (2022).
5. Aich, A. : Elastic weight consolidation (EWC): Nuts and bolts. arXiv preprint arXiv:2105.04093, (2021).
6. Prechelt, L. : Early stopping-but when?. In Neural Networks: Tricks of the trade (pp. 55-69). Berlin, Heidelberg: Springer Berlin Heidelberg, (2002).
7. Chatterji, N. S., Neyshabur, B., & Sedghi, H. : The intriguing role of module criticality in the generalization of deep networks. arXiv preprint arXiv:1912.00528, (2019).
8. Bae, W., Ren, Y., Ahmed, M. O., Tung, F., Sutherland, D. J., & Oliveira, G. L. : AdaFlood: Adaptive Flood Regularization. arXiv preprint arXiv:2311.02891, (2023).